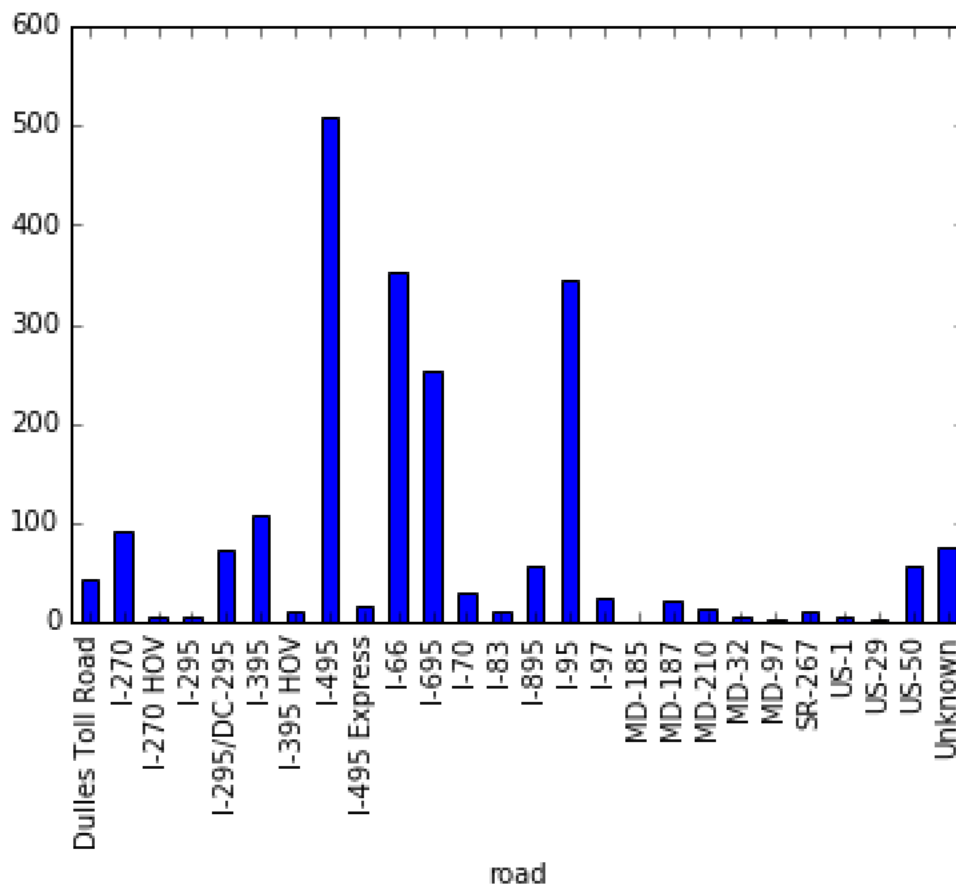


Task 4-a

Visualization

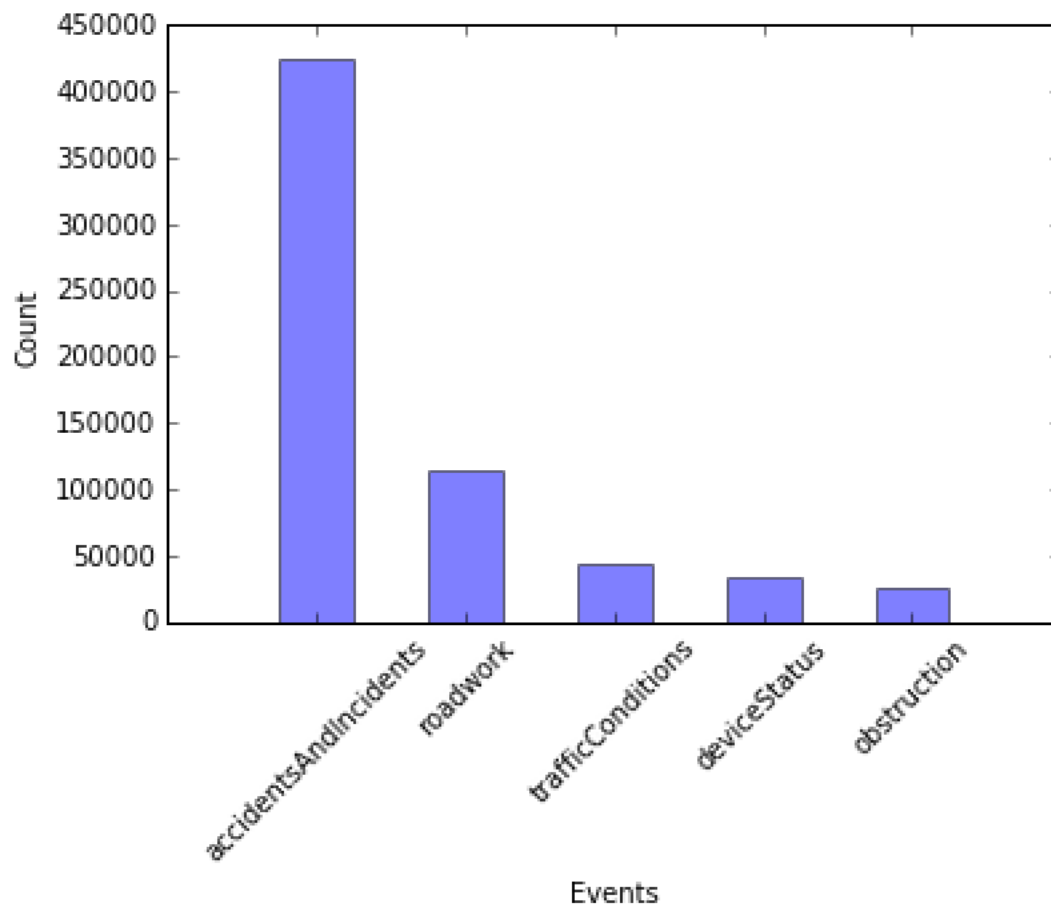


Code

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Thu Nov 03 23:08:57 2016
4
5 @author: harsh
6 """
7 import pandas as pd
8
9 df = pd.DataFrame.from_csv("detector_lane_inventory.tsv", sep="\t")
10
11 df.groupby('road').size().plot(kind='bar')
```

The visualization plots the roads vs the number of detectors on that specific road.

Task 4-b Visualization



The visualization that we have made shows the top 5 events in the events_train_holdout.csv dataset. After looking at the visualization we can say that maximum number of events are caused by accidents and incidents followed by roadwork, traffic conditions etc. The code is displayed below.

Task 4-b Code

```
# -*- coding: utf-8 -*-
"""
Created on Thu Nov 03 18:57:47 2016

@author: harsh
"""

from pandas import read_csv
import matplotlib.pyplot as plt
from pandas import DataFrame
import numpy as np

df2 = read_csv("events_train_holdout.tsv",sep="\t", error_bad_lines=False)

xy = df2['event_id'].groupby(df2['event_type'])
list1=[]
list2=[]

for name,group in xy:
    list1.append(name)
    list2.append(group.count())

df_test = DataFrame({'event_type':list1,'count':list2})
df_test=df_test.sort_values('count',ascending=False)

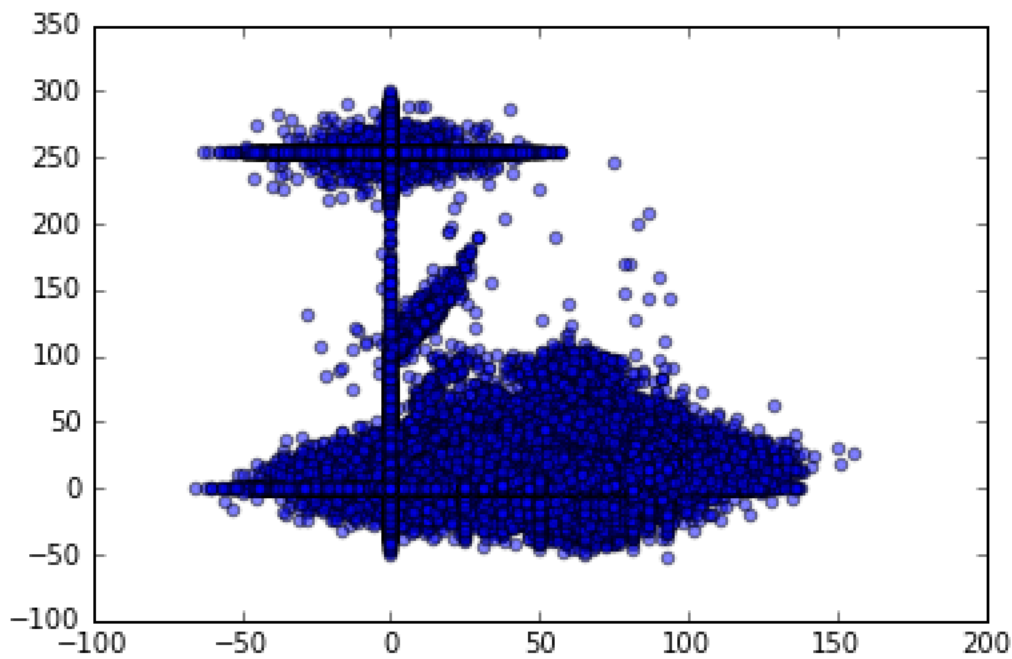
plt.bar(np.arange(5),df_test['count'][:5],align='center',width=0.5,alpha=0.5)
plt.xticks(np.arange(5),df_test['event_type'][:5],rotation=45)
plt.xlabel('Events')
plt.ylabel('Count')
plt.show()
```

Task 4c

4 c Code

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Fri Nov 04 22:41:43 2016
4
5 @author: harsh
6 """
7
8 from pandas import read_csv
9 import matplotlib.pyplot as plt
10
11 df_x = read_csv("cleaning_test_06_09.tsv", sep="\t", error_bad_lines=False)
12
13 plt.scatter(df_x['speed'], df_x['flow'], alpha=0.5)
14 plt.show()
```

4c Visualization



We have plotted speed vs flow and we see that there is a lot of dirty data as there should be a strict correlation but in this figure there is not.