San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

San Jose Traffic Accidents Prediction

Ping Chen, Mandy Wong, Nihanjali Mallavarapu, Dhruwaksh Dave

Intermediate Status Report

1. Progress towards the goal achieved so far

In this project, we attempt to predict San Jose traffic accident based on historical data acquired in City of San Jose DOT. We have collected 4 datasets and completed most of the data preparation.

weather.csv: San Jose Weather Data from 2000 to current includes the precipitation and temperature,etc.
This dataset has been cleaned and it is ready to combine with the traffic crash datasets, so that we can see how the weather condition relates to the probability of traffic accident.

traffic_volume.csv: This data including average daily traffic volume in every intersection of San Jose City. We will analyze the relationship between traffic accidents and traffic volume.

speed_survey.csv: This data contains the basic road information in San Jose, including the road length, speed limit, average traffic speed, the number of historical accident record, etc.

crash_data: This data set shows the location of individual crashes where one or more fatalities and/or severe injuries occurred during the five-year period of 2013 to 2017. The data including 940 crash data. Since these crashes are mapped to the nearest intersections, each crash contains the following information: Crash Location/Intersection, Date/Time, Injured Party, and Number of Fatal or Severe Injuries per crash.

2. Data Preparation and Visualization

Weather data preparation (Done by Mandy):

The weather data is specified in San Jose, CA from January 1st 2000 to October 20th 2019 including the date, weather record time, the precipitation, the highest temperature of the day, the lowest temperature of the day and the observed temperature in record time.

I have deleted the columns that all cells are null. Also, I have noticed that some of the data in observed temperature columns are missing, so I have used the highest temperature and the lowest temperature of the day to calculate the average temperature to fill in the missing cell in the observed temperature columns.

We will mainly use the precipitation and the temperature from this weather dataset to be considered as factors in the prediction model.

Firstly, we will use the temperature and precipitation data in this dataset to do the data integration, which is combining with the date of the traffic accidents happened in the crash dataset and finding the correlation between the weather and the possibility of traffic accidents happened in the prediction model .

Secondly, I have generated the scatter plot from 2013 to 2019 since our crash data started from 2013. We can notice that there is a certain pattern of the temperature all over the year. Based on the climate pattern shown above, we can also use this pattern correlated to the car accident data to predict the possibility of traffic accident happened in months.

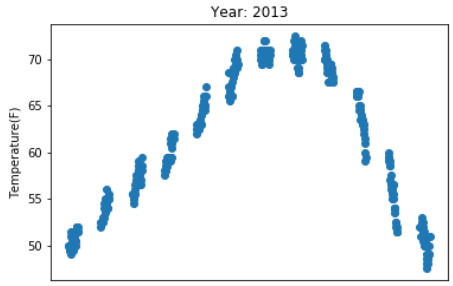
San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

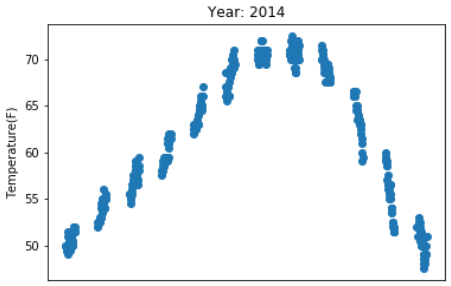
The following is some data visualization based on the weather datasets:

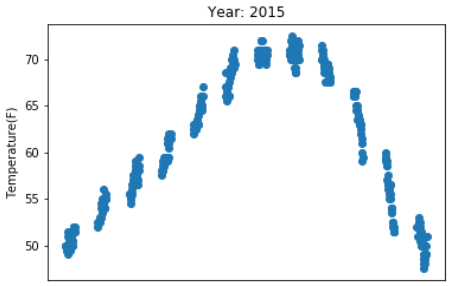
	Date	Temperature
0	20130101.0	50.0
1	20130102.0	50.0
2	20130103.0	49.5
3	20130104.0	50.0
4	20130105.0	49.5
..	...	...
360	20131227.0	49.0
361	20131228.0	49.0
362	20131229.0	49.0
363	20131230.0	50.0
364	20131231.0	51.0

	Date	Temperature
0	20140101.0	50.0
1	20140102.0	50.0
2	20140103.0	49.5
3	20140104.0	50.0
4	20140105.0	49.5
..	...	...
360	20141227.0	49.0
361	20141228.0	49.0
362	20141229.0	49.0
363	20141230.0	50.0
364	20141231.0	51.0

	Date	Temperature
0	20150101.0	50.0
1	20150102.0	50.0
2	20150103.0	49.5
3	20150104.0	50.0
4	20150105.0	49.5
..	...	...
360	20151227.0	49.0
361	20151228.0	49.0
362	20151229.0	49.0
363	20151230.0	50.0
364	20151231.0	51.0



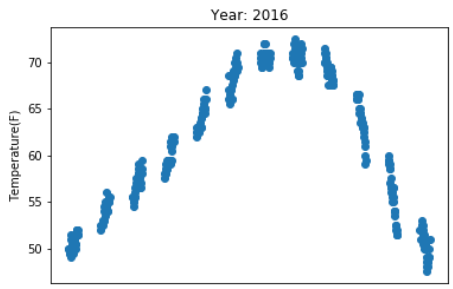


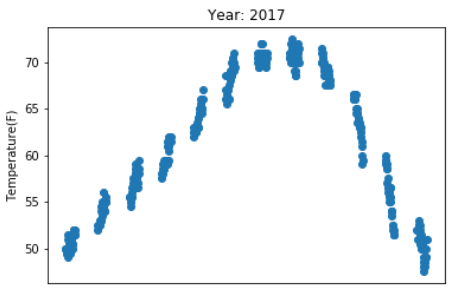


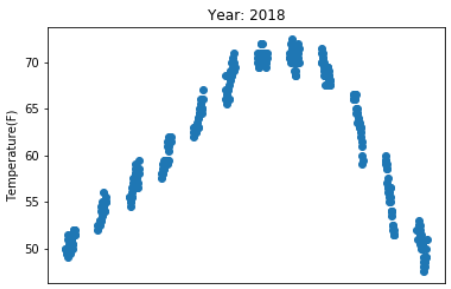
	Date	Temperature
0	20160101.0	50.0
1	20160102.0	50.0
2	20160103.0	49.5
3	20160104.0	50.0
4	20160105.0	49.5
..	...	...
361	20161227.0	49.0
362	20161228.0	49.0
363	20161229.0	49.0
364	20161230.0	50.0
365	20161231.0	51.0

	Date	Temperature
0	20170101.0	50.0
1	20170102.0	50.0
2	20170103.0	49.5
3	20170104.0	50.0
4	20170105.0	49.5
..	...	...
360	20171227.0	49.0
361	20171228.0	49.0
362	20171229.0	49.0
363	20171230.0	50.0
364	20171231.0	51.0

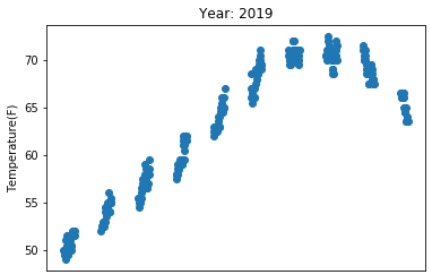
	Date	Temperature
0	20180101.0	50.0
1	20180102.0	50.0
2	20180103.0	49.5
3	20180104.0	50.0
4	20180105.0	49.5
..	...	...
360	20181227.0	49.0
361	20181228.0	49.0
362	20181229.0	49.0
363	20181230.0	50.0
364	20181231.0	51.0







	Date	Temperature
0	20190101.0	50.0
1	20190102.0	50.0
2	20190103.0	49.5
3	20190104.0	50.0
4	20190105.0	49.5
..	...	...
288	20191016.0	63.5
289	20191017.0	64.0
290	20191018.0	63.5
291	20191019.0	63.5
292	20191020.0	63.5



San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

Road Network Data Preparation(Done by Ping):

I collected road datasets from San Jose DOT with basic information in San Jose. I referred to two of the road datasets: one is average daily traffic data and the other is speed limit and speed survey data. Both the two csv dataset are based on city of San Jose streets.

(1) Data preprocessing:

a. Dimension Reduction:

First, delete irrelevant columns, such as Last Edit person, global ID, data source, etc. After dimension reduction, the size of average daily traffic data reduced from (2044, 23) to (2044, 7) and the size of average daily traffic data reduced from (580, 27) to (580, 10). Below shows the features after reduction in these two datasets:

```
In [286]: traffic = traffic.drop(["X","Y","OBJECTID","FACILITYID", "INTID", "INTID", "TRAVELDIRE","COUNTDATE","CITY", "ADTONE",'
          traffic.keys()

Out[286]: Index(['LATITUDE', 'LONGITUDE', 'ADT', 'STREETONE', 'DIRECTION', 'STREETTWO',
                 'NEARINTERS'],
                dtype='object')
```

```
In [246]: speed = speed.drop(["OBJECTID","ROUTE","DATE","on_hold", "CD", "PD", "GlobalID","reason_for_on_hold", "Comment", "Globe
          speed.keys()

Out[246]: Index(['STREET', 'START', 'END_', 'SPD', 'LEN', 'ACC', 'VOL', 'F85th', 'LOCAL',
                 'F50th'],
                dtype='object')
```

b. Handle Missing data:

(1) Average daily traffic data:

Through a quick description of the data, I found that the feature ADT(average daily traffic data) has 9 missing data. I fill in these NA with median value of this feature.

```
traffic.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2044 entries, 0 to 2043
Data columns (total 7 columns):
LATITUDE      2044 non-null float64
LONGITUDE     2044 non-null float64
ADT           2035 non-null float64
STREETONE     2044 non-null object
DIRECTION     2044 non-null object
STREETTWO     2044 non-null object
NEARINTERS    2044 non-null object
dtypes: float64(3), object(4)
memory usage: 111.9+ KB
```

(2) Speed survey data:

In speed survey data, it lacks the important location values about latitude and longitude of the intersection in each item, which is useful in data integration with crash data. I used googlemap geocode API to find the latitude and longitude for each item. Below are the first 10 lines of this data after adding latitude and longitude column:

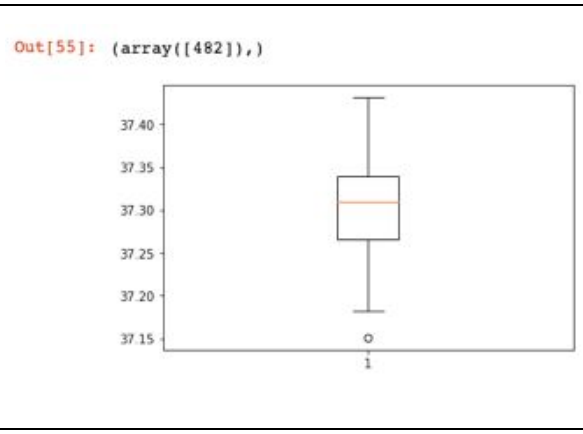
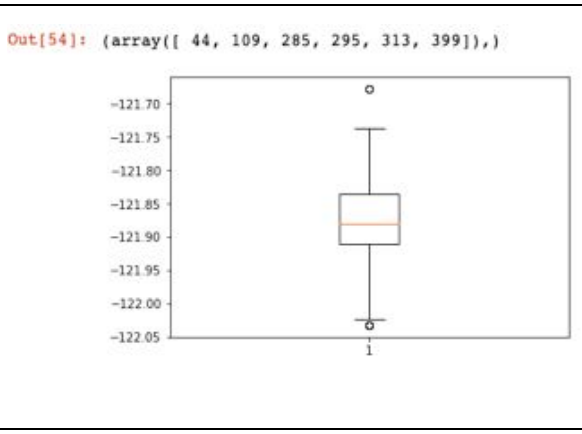
San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

```
speed.head(10)
```

Out[6]:

	STREET	START	END_	SPD	LEN	ACC	VOL	F85th	LOCAL	F50th	lat	lng
0	AIRPORT BL	COLEMAN AV	SKYPORT DR	35	7200.0	7.0	14450.0	41.0	0	37.0	37.3507	-121.921
1	AIRPORT PW	AIRPORT BL	MATRIX BL	35	2500.0	13.0	18350.0	39.0	0	35.0	37.3639	-121.929
2	ALLEN AV	SANTA TERESA BL	BLOSSOM HILL RD	30	3700.0	5.0	3500.0	32.5	0	28.5	37.2405	-121.858
3	ALMA AV	FIRST ST	SENTER RD	35	3970.0	20.0	9300.0	37.5	0	34.0	37.3164	-121.874
4	ALMA AV	VINE ST	FIRST ST	30	2000.0	47.0	17000.0	33.0	0	30.0	37.3128	-121.879
5	ALMA AV	MINNESOTA ST	VINE ST	35	2950.0	29.0	13200.0	37.5	0	34.5	37.3089	-121.888
6	ALMADEN AV	ALMA AV	REED ST	30	5000.0	50.0	7900.0	33.3	0	30.0	37.3135	-121.878
7	ALMADEN BL	100' N/O GRANT ST	SANTA CLARA ST	30	4300.0	22.0	13100.0	33.5	0	29.0	37.3323	-121.894
8	ALMADEN EX	SR-87	SAN JOSE AV	45	2200.0	9.0	13500.0	48.0	0	44.0	37.189	-121.845
9	ALMADEN EX	SAN JOSE AV	ALMA AV	40	2100.0	33.0	20200.0	42.0	0	38.0	37.3074	-121.878

Plot the boxplot of latitude and attitude to make sure that the API return the correct values. There is one outlier for latitude and 6 outliers for longitude. I have checked that the corresponding items located in the boundaries of San Jose and the latitude and longitude values are correct.

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>Latitude boxplot</td>
<td>Longitude boxplot</td>
</tr>
</table>

Then getting insight about speed survey data and there are some missing data in features of F85th, F50, road length, ACC(the number of historical traffic accident). The F85th means the speed at or below which 85 percent of all vehicles are observed to travel under free-flowing conditions past a monitored point and the F50 is 50 percent. So I fill in F85th NA with 1.1 times of corresponding limited speed and F50 with the limited speed, which seems more reasonable according to the driving style in San Jose. There are 6 missing values in ACC and road length, I dropped these 6 rows since I have no idea about these values.

Below is the summary of speed survey data after handling missing data:

```
speed.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 574 entries, 0 to 579
Data columns (total 16 columns):
STREET          574 non-null object
START           574 non-null object
END_            574 non-null object
SPD             574 non-null int64
LEN             574 non-null float64
ACC             574 non-null float64
VOL             574 non-null float64
F85th           574 non-null float64
LOCAL           574 non-null int64
F50th           574 non-null float64
lat             574 non-null float64
lng             574 non-null float64
```

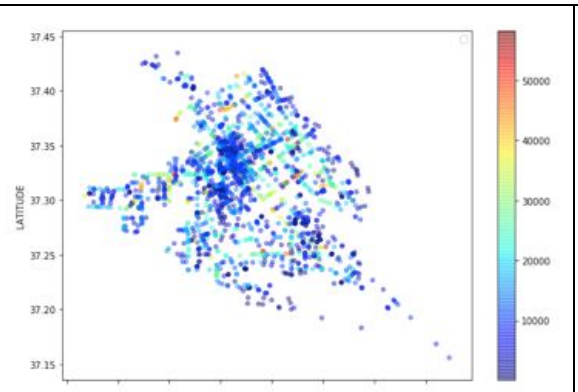
San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

c. Data Visualization:

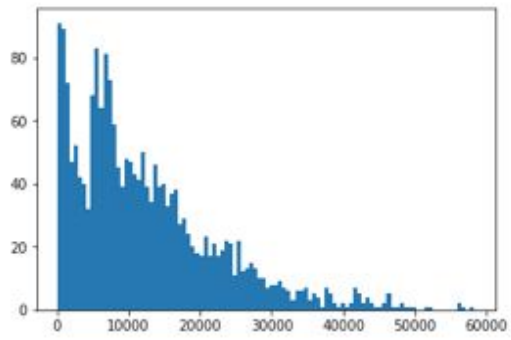
(1) Average daily traffic data:

Create a scatter plot to visualize traffic volume with geographical information. The pattern looks exactly like San Jose. I used a color map which ranges from blue (low values) to red(high values). The high-density areas are approximately around downtown, which has a lot of intersections. The pointer with lighter colors are located on the main road or near the highway entrance.

A histogram below shows the number of instances that have a traffic volume at given value range. The max traffic volumes is 58274.0 and min value is 100, average value is 12365.



A geographical scatter plot of the data



Histogram for ADT

Arrange the data by traffic volume we can see the top 10 items. It is reasonable that the main road like Tully RD, Capitol Express and Blossom Hill RD have the highest daily traffic volume.

```
traffic.sort_values(by=['ADT'],ascending = False).head(10)
```

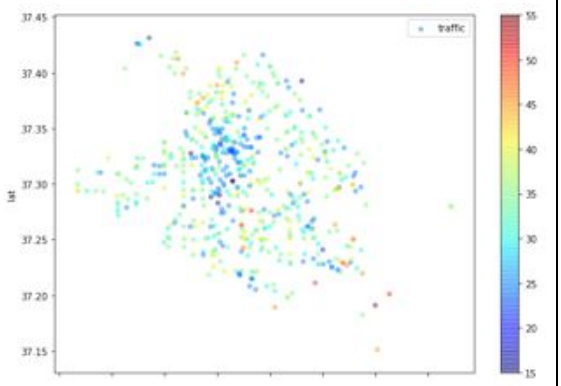
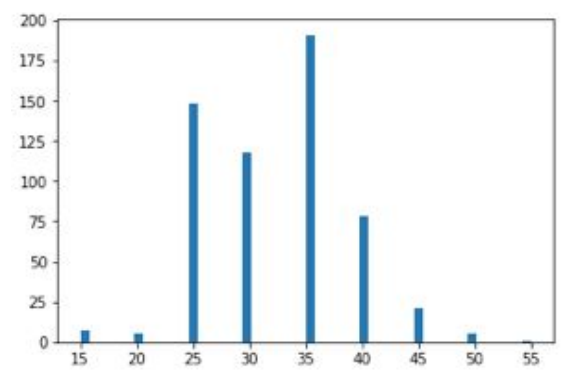
Out[311]:

	LATITUDE	LONGITUDE	ADT	STREETONE	DIRECTION	STREETTWO	NEARINTERS
221	37.315811	-121.835566	58274.0	Tully Rd	E of	McLaughlin Av	Tully Rd & McLaughlin Av
1336	37.355506	-121.834558	57107.0	CAPITOL EXPY	E of	MASSAR	CAPITOL EXPY&MASSAR
276	37.320453	-121.828685	56344.0	Tully Rd	E of	Alvin Av	Tully Rd & Alvin Av
401	37.257227	-121.797167	56000.0	Blossom Hill Rd	E of	Hwy-101	Blossom Hill Rd & Hwy-101
219	37.323364	-121.945314	51883.0	Stevens Creek Bl	E of	Baywood Av	Stevens Creek Bl & Baywood Av
973	37.322281	-121.826050	51822.0	Tully Rd	W of	King Rd	Tully Rd & King Rd
1029	37.382344	-121.900446	49988.0	BROKAW RD	E of	RIDDER PARK DR	BROKAW RD & RIDDER PARK DR
971	37.314997	-121.836772	49545.0	Tully Rd	W of	McLaughlin Av	Tully Rd & McLaughlin Av
280	37.313068	-121.839772	48908.0	Tully Rd	E of	Lucretia Av	Tully Rd & Lucretia Av
951	37.365579	-121.849631	48289.0	McKee Rd	E of	Jackson Av	McKee Rd & Jackson Av

(2) Average daily traffic data:

Create a scatter plot to visualize speed limit with geographical information and a histogram to see the number of instances locate in a given range.

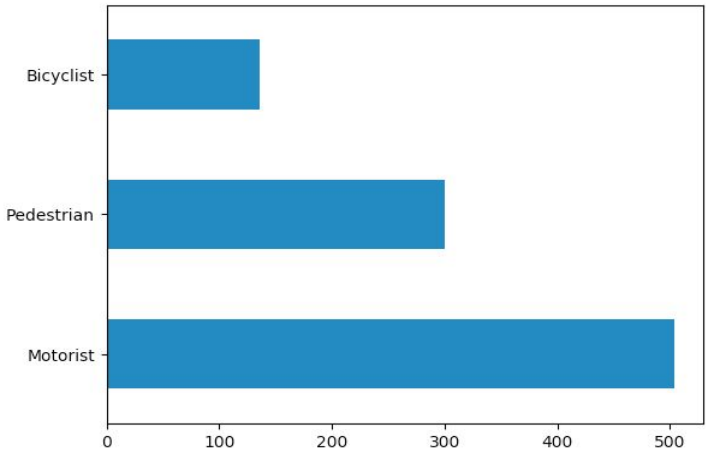
San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

	
A geographical scatter plot of the data	Histogram for SPD

Motor vehicle crash data preparation: (Done by Nihanjali and Dhruwaksh)

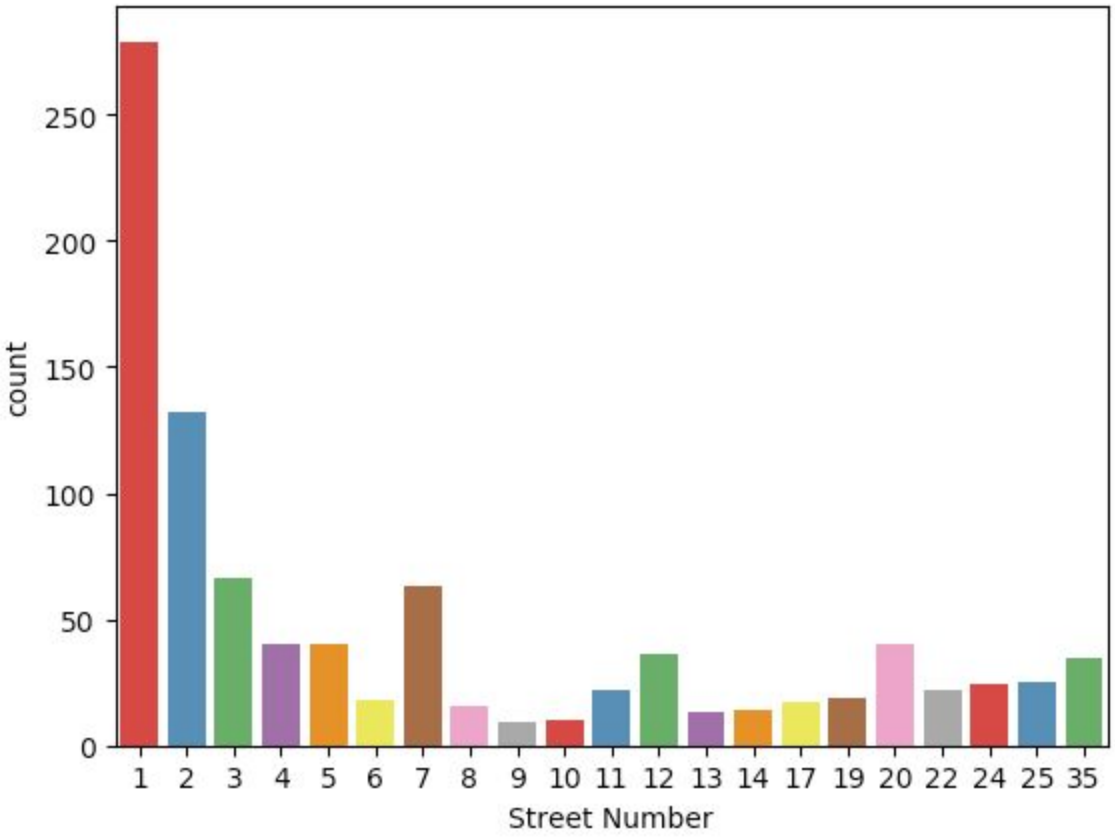
A. Data Visualization :

1. The plot shows the frequency of actors involved in the accidents. I.e. Pedestrians, bicyclists and motorists.



2. This plot shows the accident count on the streets. When we labelled the street names, then the individual names overlapped on x axis. So we gave the street names a number and plotted the count against that.

San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report



B. Data Preprocessing

Obtained from a dataset having records for all the crashes in each of the 50 states of The United States. We narrowed down the data on the basis of location ie., City and selected the records pertinent to San Jose. Preprocessing steps like removal of the irrelevant data, reduction of dimensions, filling in placeholders for null fields was done on the dataset.
The data has been cleaned and is ready for integrating with the other datasets on the basis of select attributes. This will help us have a uniform dataset to run machine learning algorithms on so as to achieve efficient and accurate predictions.
We'll have to apply feature scaling for standardizing the representations of our data so as to make integrating with the other datasets convenient.

Data integration (Partially complete) (Done by Ping)

We need to integrate Road network data and weather data into traffic accident dataset. Currently I have finished the integration of Road network data and traffic accident data. I will complete the integration of weather data in our next stage.

In data integration, I matched attribute from one database to another. Below is a brief summary of the integrated dataset:

San Jose State University
Fall 2019
CMPE 255 Sec 02
Group-6 Intermediate Status Report

```
dataset.info()
dataset.shape
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 853 entries, 0 to 939
Data columns (total 16 columns):
LONGITUDE              853 non-null float64
LATITUDE               853 non-null float64
Date                   853 non-null object
AStreet                853 non-null object
BStreet                853 non-null object
FatalInjuries          853 non-null object
MajorInjuries          853 non-null object
Involving              853 non-null object
Nearest_Intersection   853 non-null object
SPD                    853 non-null object
LEN                    853 non-null float64
ACC                    853 non-null float64
VOL                    853 non-null float64
F85th                  853 non-null float64
LOCAL                  853 non-null object
F50th                  853 non-null float64
dtypes: float64(7), object(9)
memory usage: 113.3+ KB
```

```
Out[75]: (853, 16)
```

3. Difficulties being encountered and how you resolve them (Done by Ping)

The most difficult is to combine the datasets, our goal is to combine all datasets into one. I put in a lot of effort to combine traffic_volume.csv, speed_survey.csv and traffic_accident.csv. The most difficult is how match the attribute in one dataset to another dataset and how to find the correct corresponding data tuple in another dataset. For example, in the traffic accident csv and speed survey csv, the street names are represented by upper case, but in traffic volume csv the street names are lowercase. Then in each item in traffic accident csv, use the intersection name to find the corresponding item in the other two files. If the intersection name cannot find exactly in other files, I will check if other intersections in the same street can be found in the files. If so, I choose the nearest intersection to the intersection in traffic accident file by computing the minimum distance based on their latitude and longitude. Finally, I integrate 853 items out of 940 items. For the remaining items, Since I cannot find any relative information in other two files, I dropped them temporarily.

4. Remaining tasks

The next step we need to do is finishing the data integration. Then we need to get insight into the complete dataset to find the correlation between. Another thing is since all the dataset we found till now is positive dataset, we need to create negative data.
After all data sets are ready, we will implement the assigned Machine Learning Algorithm with our training data.
We will compare the accuracy of the model after implemented and improve our model in order to obtain the most accurate prediction of traffic accidents in San Jose.

5. Any others that you think is relevant

Team Roles:
Data Preparation: Since we have three data resources, each of us will choose one feature to do data preparation and the rest of us will do data integration.
Algorithms: Each of us will choose one algorithm to train the data and then compare the results of the four mode:

Name	Data Preparation	Data cleaned?	Machine Learning Algorithms
Ping Chen	Road Network Dataset Data Integration	Y	ANN
Mandy Wong	Climate Data	Y	SVM
Nihanjali Mallavarapu	Motor Vehicle Crash Data		Logistic Regression
Dhruwaksh Dave	Motor Vehicle Crash Data		Random forests