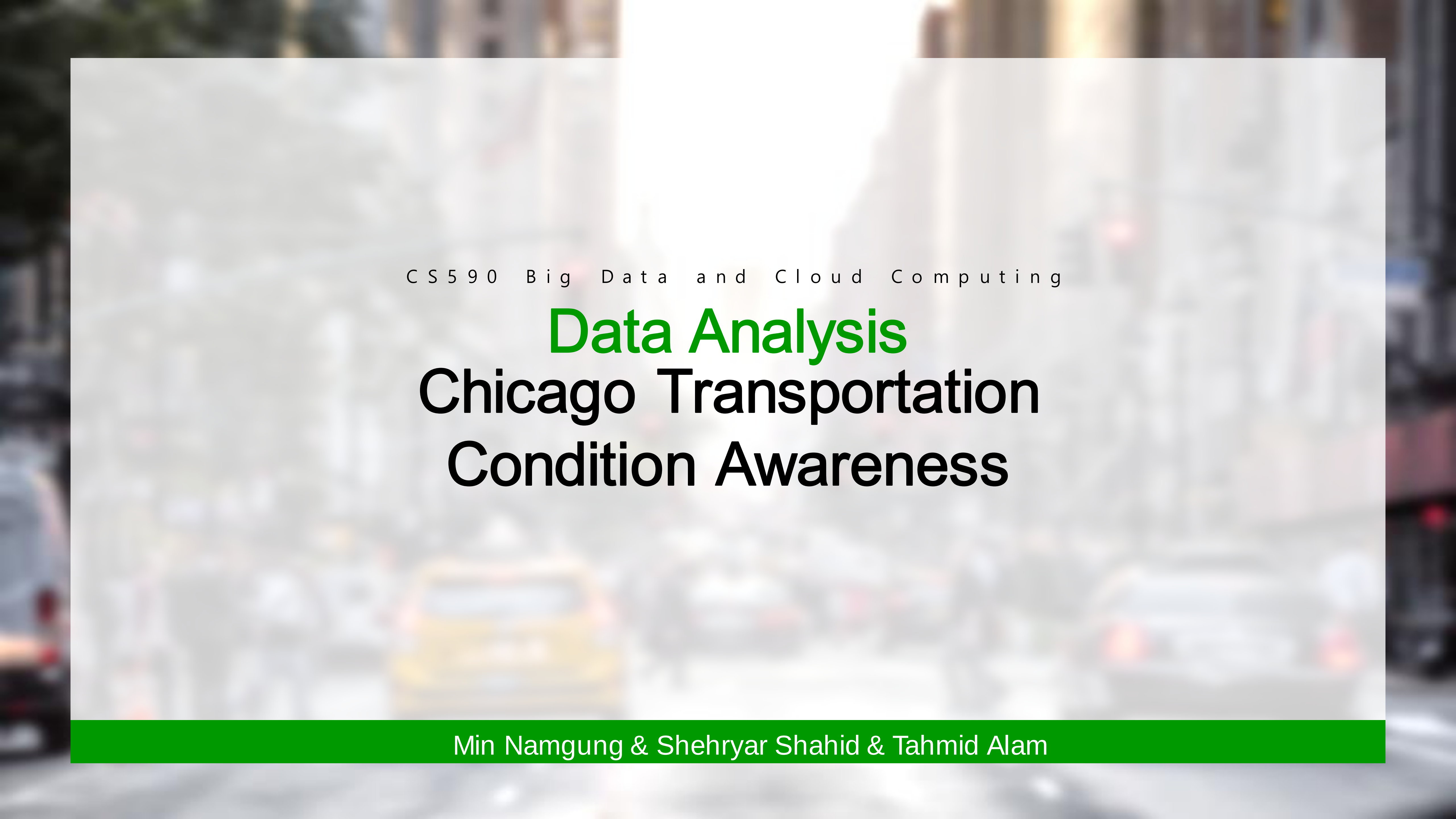


CS590 Big Data and Cloud Computing

Data Analysis
Chicago Transportation Condition Awareness

Min Namgung & Shehryar Shahid & Tahmid Alam

1

Overview of project

- The project problem description
- Applications
- Motivation of project problem
- The goal and objectives

2

Approach & Implementation

- Challenges
- Methodology
- General methods and technologies
- Approaches

3

Interpretation & Result

- Interpretation of results
- Results
- Recommendation

4

Summary & Conclusion

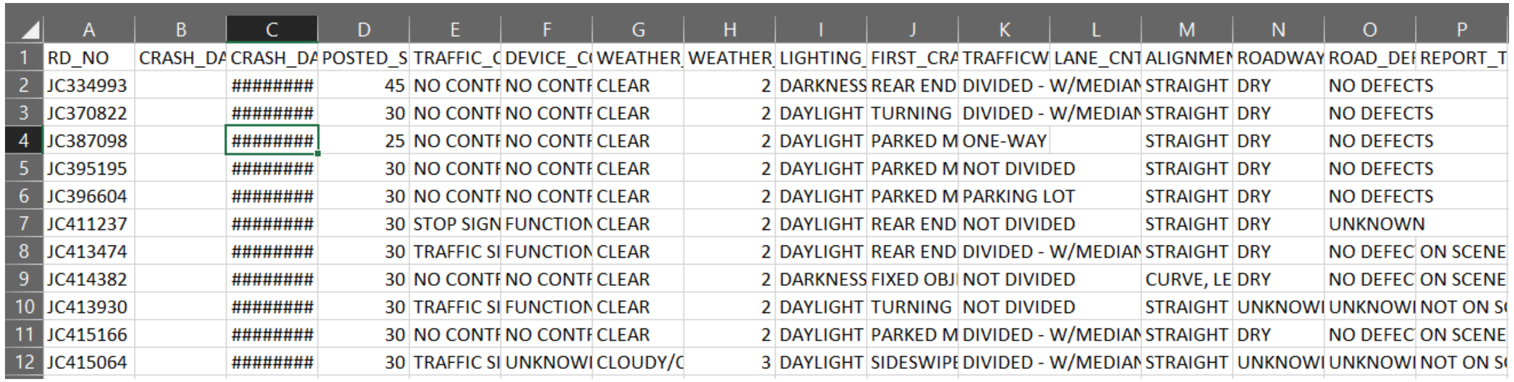
- A summary of project
- A description of limitations of current work
- A description of possible future work directions

Overview of project

Tools and dataset we used for the project.

Chicago car crash data

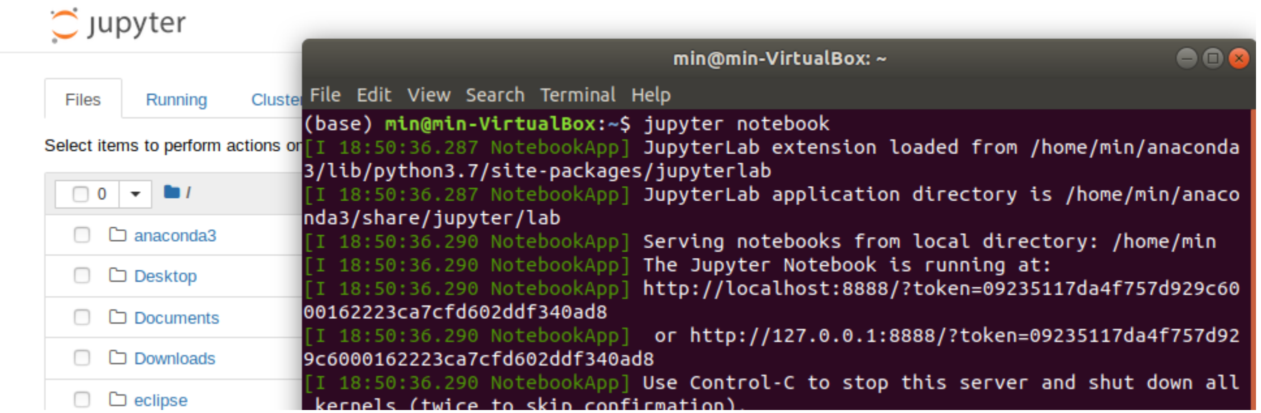
Public open dataset, big data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	RD_NO	CRASH_DA	CRASH_DA	POSTED_S	TRAFFIC_C	DEVICE_C	WEATHER_	WEATHER_	LIGHTING_	FIRST_CRA	TRAFFICW	LANE_CNT	ALIGNMEN	ROADWAY	ROAD_DEF	REPORT_T
2	JC334993		########	45	NO CONTR	NO CONTR	CLEAR	2	DARKNESS	REAR END	DIVIDED - W/MEDIAN		STRAIGHT	DRY	NO DEFECTS	
3	JC370822		########	30	NO CONTR	NO CONTR	CLEAR	2	DAYLIGHT	TURNING	DIVIDED - W/MEDIAN		STRAIGHT	DRY	NO DEFECTS	
4	JC387098		########	25	NO CONTR	NO CONTR	CLEAR	2	DAYLIGHT	PARKED M	ONE-WAY		STRAIGHT	DRY	NO DEFECTS	
5	JC395195		########	30	NO CONTR	NO CONTR	CLEAR	2	DAYLIGHT	PARKED M	NOT DIVIDED		STRAIGHT	DRY	NO DEFECTS	
6	JC396604		########	30	NO CONTR	NO CONTR	CLEAR	2	DAYLIGHT	PARKED M	PARKING LOT		STRAIGHT	DRY	NO DEFECTS	
7	JC411237		########	30	STOP SIGN	FUNCTION	CLEAR	2	DAYLIGHT	REAR END	NOT DIVIDED		STRAIGHT	DRY	UNKNOWN	
8	JC413474		########	30	TRAFFIC SI	FUNCTION	CLEAR	2	DAYLIGHT	REAR END	DIVIDED - W/MEDIAN		STRAIGHT	DRY	NO DEFEC	ON SCENE
9	JC414382		########	30	NO CONTR	NO CONTR	CLEAR	2	DARKNESS	FIXED OBJ	NOT DIVIDED		CURVE, LE	DRY	NO DEFEC	ON SCENE
10	JC413930		########	30	TRAFFIC SI	FUNCTION	CLEAR	2	DAYLIGHT	TURNING	NOT DIVIDED		STRAIGHT	UNKNOWN	UNKNOWN	NOT ON S
11	JC415166		########	30	NO CONTR	NO CONTR	CLEAR	2	DAYLIGHT	PARKED M	DIVIDED - W/MEDIAN		STRAIGHT	DRY	NO DEFEC	ON SCENE
12	JC415064		########	30	TRAFFIC SI	UNKNOWN	CLOUDY/C	3	DAYLIGHT	SIDESWIPE	DIVIDED - W/MEDIAN		STRAIGHT	UNKNOWN	UNKNOWN	NOT ON S

Tools to use

Jupyter notebook – python && Pyspark



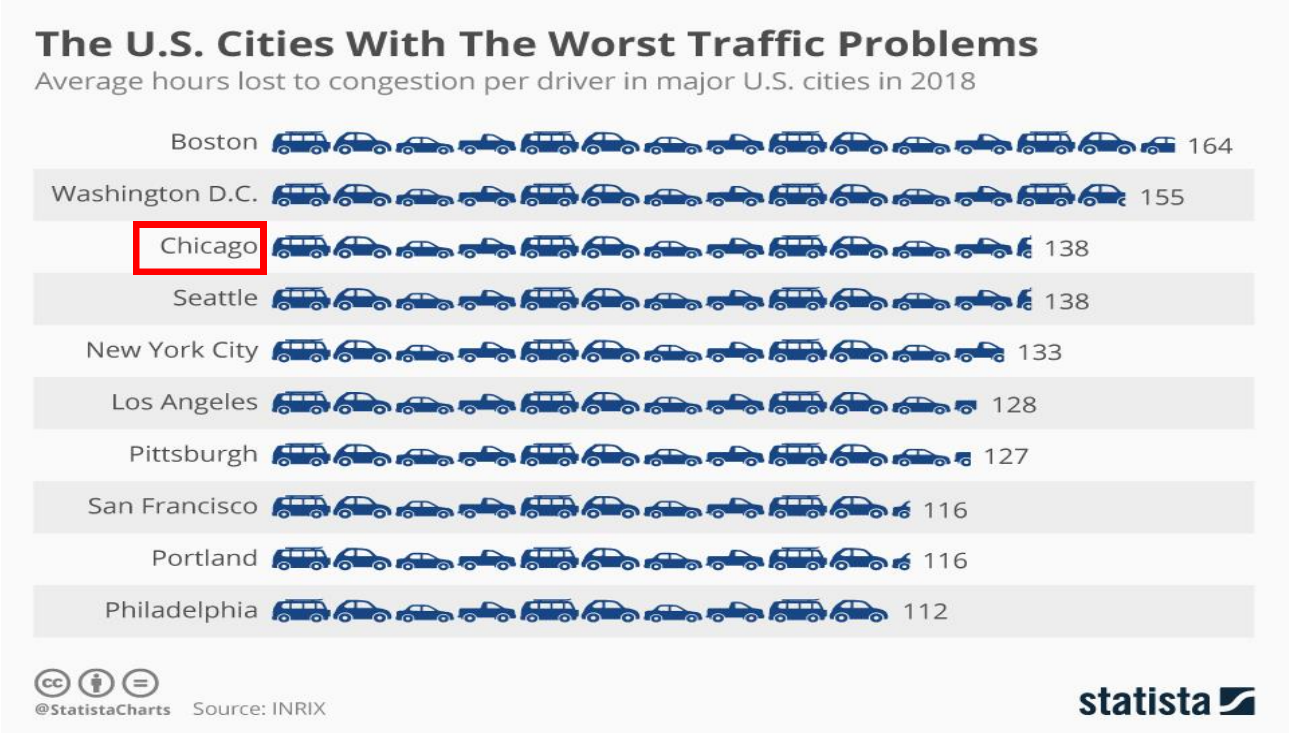
```
(base) min@min-VirtualBox:~$ jupyter notebook
[I 18:50:36.287 NotebookApp] JupyterLab extension loaded from /home/min/anaconda
3/lib/python3.7/site-packages/jupyterlab
[I 18:50:36.287 NotebookApp] JupyterLab application directory is /home/min/anaco
nda3/share/jupyter/lab
[I 18:50:36.290 NotebookApp] Serving notebooks from local directory: /home/min
[I 18:50:36.290 NotebookApp] The Jupyter Notebook is running at:
[I 18:50:36.290 NotebookApp] http://localhost:8888/?token=09235117da4f757d929c60
00162223ca7cfd602ddf340ad8
[I 18:50:36.290 NotebookApp]  or http://127.0.0.1:8888/?token=09235117da4f757d92
9c6000162223ca7cfd602ddf340ad8
[I 18:50:36.290 NotebookApp] Use Control-C to stop this server and shut down all
 kernels (twice to skip confirmation).
```

1.1

Overview of project

■ Problem statement

Busy Chicago's traffic leads to many problems such as **injuries** or **car crashes** which cause another collision. The car crash can cause other **traffic jams** and it brings about **more busy traffic** which possibly generates another car accident



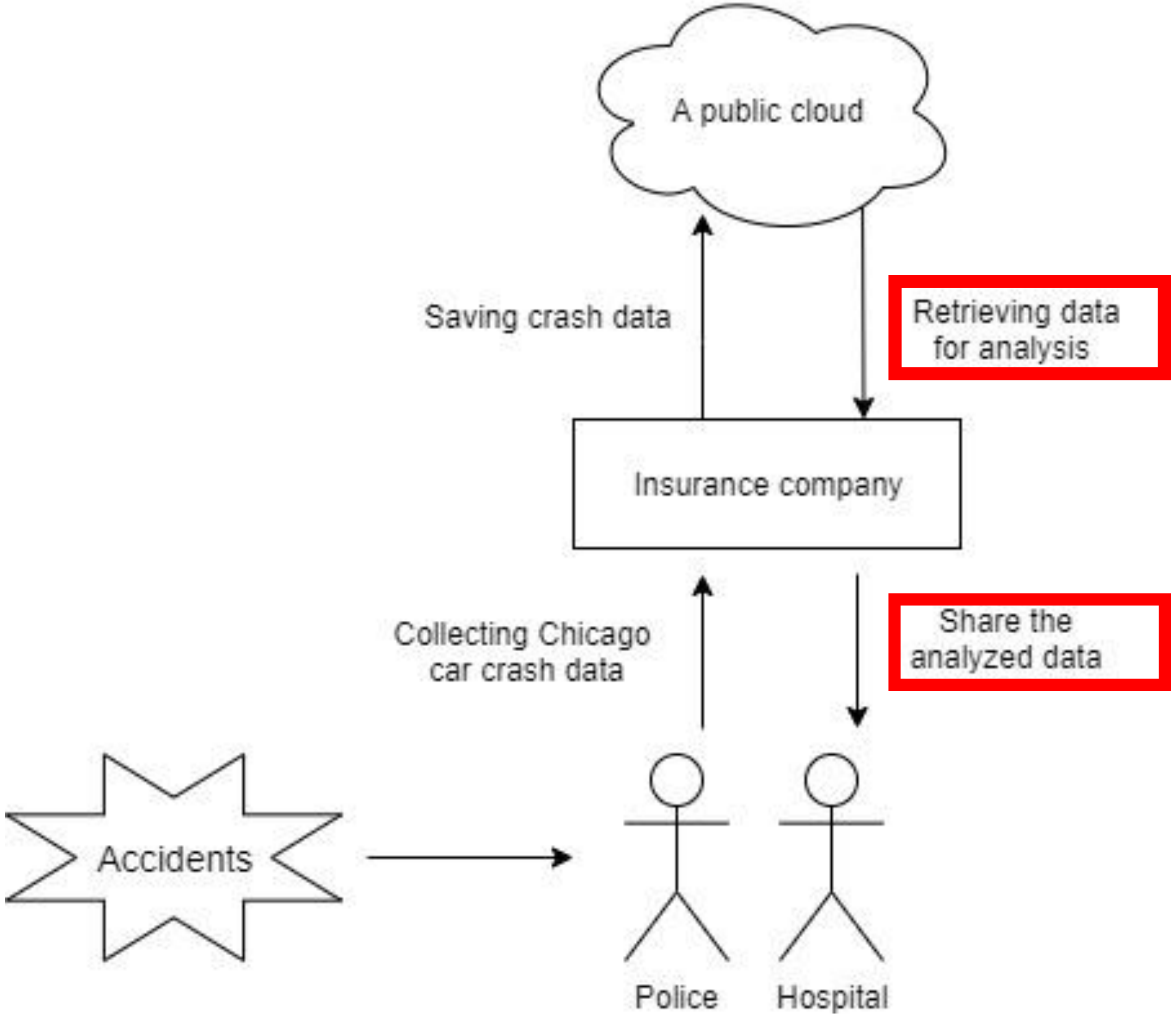
The U.S. Cities With The Worst Traffic Problems

Average hours lost to congestion per driver in major U.S. cities in 2018

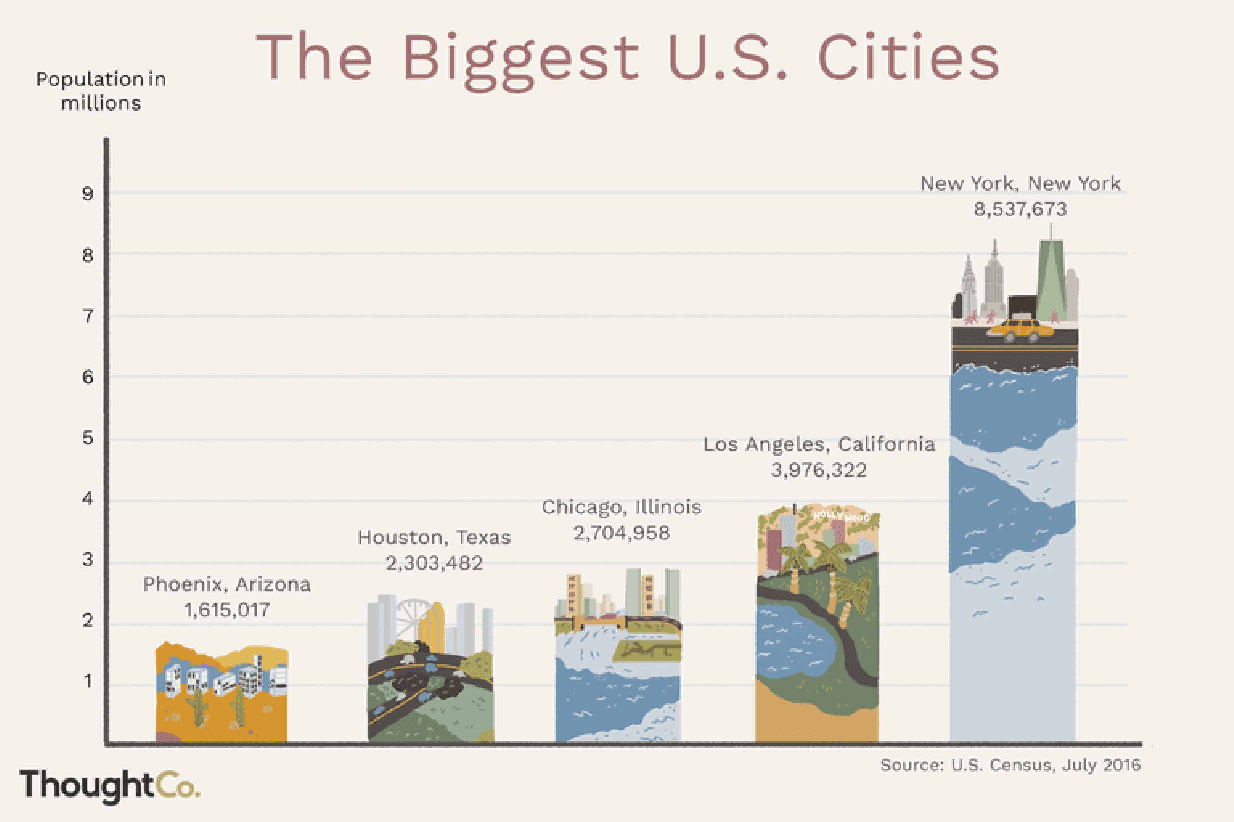
Source: INRIX

1.2
Overview of project

▪ Application



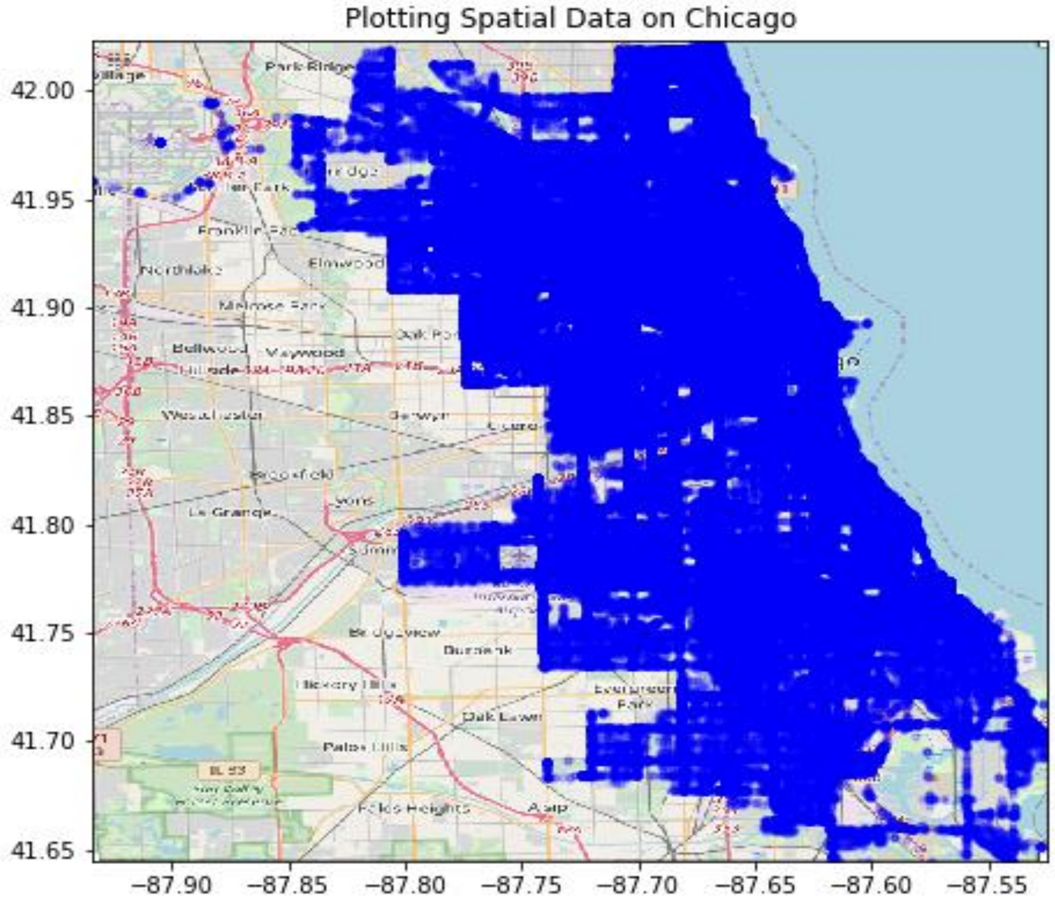
▪ Motivation of project



- The third-largest city in U.S.
- A large amount of daily commuters and population
- Easy access to public data
- Car crash data has collected for th e past 6 years -> a huge data colle ction

1.3
Overview of project

- Motivation of project


Plotting Spatial Data on Chicago

1.4

Overview of project

Goal of project

- To analyze Chicago's traffic accident data for the past 6 years **to offer valuable insight to bot h the general public, and the police**
- The car crashes dataset include some **external factors** such as weather conditions, traffic way types, road defection, or the location of accidents.
- With this additional information, **Chicago drivers** will be able to **avoid the most accident-like conditions**.
- Stakeholders such as **police or insurance company** will be able to **understand** how to help **for reducing** these situations, and **how to prevent accidents** before they occur

2.1

Approach & Implementation

■ Challenge

- Collecting the data which will help to give us better analysis result.
- Some of the important data which could have made more impact on the analysis are missing.
- Lack of domain knowledge to choose features which will give us more robust result.
- Finding some important library in spark environment to support our approach.
- Not enough compatibility with numpy library.

2.2
Approach & Implementation

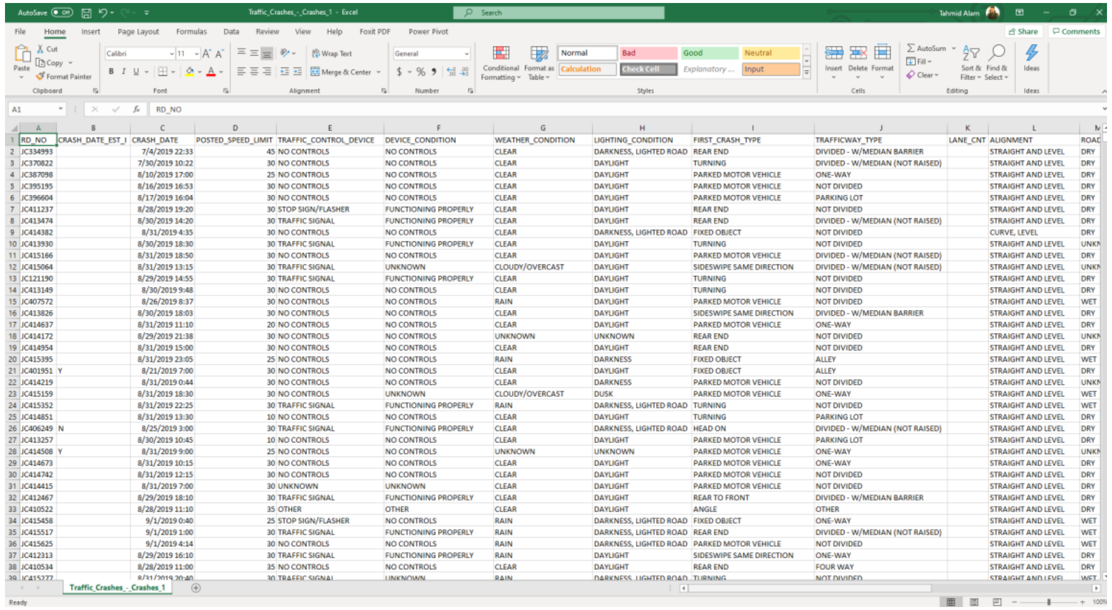
Methodology

- Collection of data
- Exploration of collected data
- Data cleaning
- Feature selection
- Elbow method
- Feature indexing
- Scaling of data
- Kmeans Algorithm

2.3
Approach & Implementation

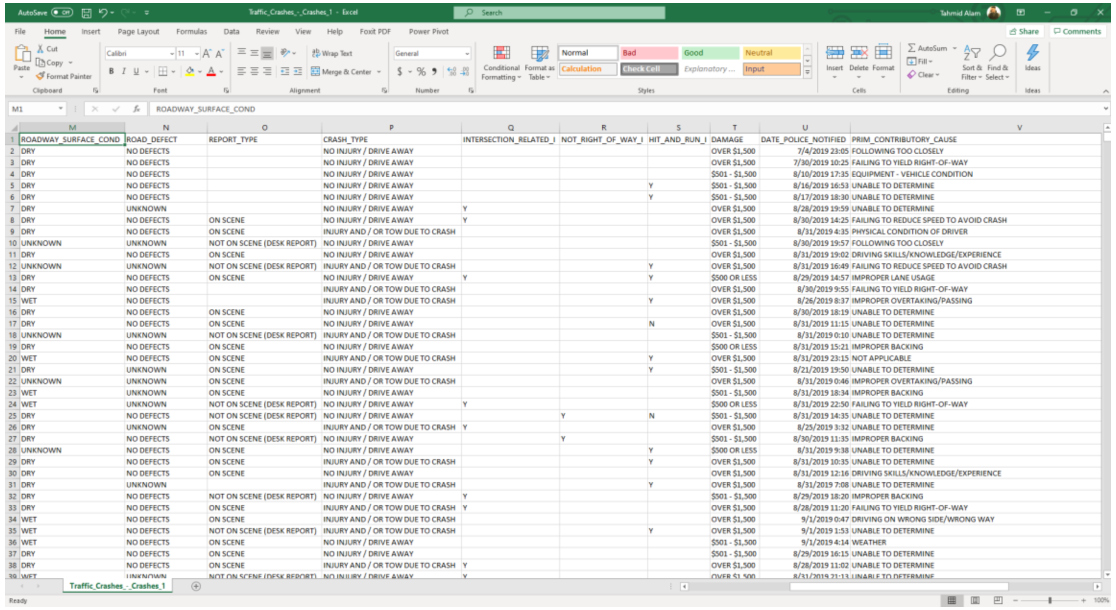
Collection of data

- We collected our data from city of Chicago
- There are plenty of car crash data to work
- There are many outcomes and analysis with our approach
- The data we collected consists of almost 350K+ rows and 48 columns.
- Data range from 2003-2019.


Traffic_Crashes_-_Crashes_1 - Excel
Tahmid Alam
Traffic_Crashes_-_Crashes_1
Ready

A1 RD_NO

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RD_NO	CRASH_DATE_EST_I	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDITION	LIGHTING_CONDITION	FIRST_CRASH_TYPE	TRAFFICWAY_TYPE	LANE_CNT	ALIGNMENT	ROAD
2	JC334993		7/4/2019 22:33	45	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	REAR END	DIVIDED - W/MEDIAN BARRIER		STRAIGHT AND LEVEL	DRY
3	JC370822		7/30/2019 10:22	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	TURNING	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	DRY
4	JC387098		8/10/2019 17:00	25	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	ONE-WAY		STRAIGHT AND LEVEL	DRY
5	JC395195		8/16/2019 16:53	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
6	JC396604		8/17/2019 16:04	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	PARKING LOT		STRAIGHT AND LEVEL	DRY
7	JC411237		8/28/2019 19:20	30	STOP SIGN/FLASHER	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	REAR END	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
8	JC413474		8/30/2019 14:20	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	REAR END	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	DRY
9	JC414382		8/31/2019 4:35	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS, LIGHTED ROAD	FIXED OBJECT	NOT DIVIDED		CURVE, LEVEL	DRY
10	JC413930		8/30/2019 18:30	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED		STRAIGHT AND LEVEL	UNKN
11	JC415166		8/31/2019 18:50	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	DRY
12	JC415064		8/31/2019 13:15	30	TRAFFIC SIGNAL	UNKNOWN	CLOUDY/OVERCAST	DAYLIGHT	SIDESWIPE SAME DIRECTION	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	UNKN
13	JC121190		8/29/2019 14:55	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
14	JC413149		8/30/2019 9:48	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	TURNING	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
15	JC407572		8/26/2019 8:37	30	NO CONTROLS	NO CONTROLS	RAIN	DAYLIGHT	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	WET
16	JC413826		8/30/2019 18:03	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	SIDESWIPE SAME DIRECTION	DIVIDED - W/MEDIAN BARRIER		STRAIGHT AND LEVEL	DRY
17	JC414637		8/31/2019 11:10	20	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	ONE-WAY		STRAIGHT AND LEVEL	DRY
18	JC414172		8/29/2019 21:38	30	NO CONTROLS	NO CONTROLS	UNKNOWN	UNKNOWN	REAR END	NOT DIVIDED		STRAIGHT AND LEVEL	UNKN
19	JC414954		8/31/2019 15:00	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	REAR END	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
20	JC415395		8/31/2019 23:05	25	NO CONTROLS	NO CONTROLS	RAIN	DARKNESS	FIXED OBJECT	ALLEY		STRAIGHT AND LEVEL	WET
21	JC401951	Y	8/21/2019 7:00	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	FIXED OBJECT	ALLEY		STRAIGHT AND LEVEL	DRY
22	JC414219		8/31/2019 0:44	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	UNKN
23	JC415159		8/31/2019 18:30	30	NO CONTROLS	UNKNOWN	CLOUDY/OVERCAST	DUSK	PARKED MOTOR VEHICLE	ONE-WAY		STRAIGHT AND LEVEL	WET
24	JC415352		8/31/2019 22:25	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	RAIN	DARKNESS, LIGHTED ROAD	TURNING	NOT DIVIDED		STRAIGHT AND LEVEL	WET
25	JC414851		8/31/2019 13:30	10	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	TURNING	PARKING LOT		STRAIGHT AND LEVEL	DRY
26	JC406249	N	8/25/2019 3:00	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DARKNESS, LIGHTED ROAD	HEAD ON	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	DRY
27	JC413257		8/30/2019 10:45	10	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	PARKING LOT		STRAIGHT AND LEVEL	DRY
28	JC414508	Y	8/31/2019 9:00	25	NO CONTROLS	NO CONTROLS	UNKNOWN	UNKNOWN	PARKED MOTOR VEHICLE	ONE-WAY		STRAIGHT AND LEVEL	UNKN
29	JC414673		8/31/2019 10:15	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	ONE-WAY		STRAIGHT AND LEVEL	DRY
30	JC414742		8/31/2019 12:15	30	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
31	JC414415		8/31/2019 7:00	30	UNKNOWN	UNKNOWN	CLEAR	DAYLIGHT	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	DRY
32	JC412467		8/29/2019 18:10	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DAYLIGHT	REAR TO FRONT	DIVIDED - W/MEDIAN BARRIER		STRAIGHT AND LEVEL	DRY
33	JC410522		8/28/2019 11:10	35	OTHER	OTHER	CLEAR	DAYLIGHT	ANGLE	OTHER		STRAIGHT AND LEVEL	DRY
34	JC415458		9/1/2019 0:40	25	STOP SIGN/FLASHER	NO CONTROLS	RAIN	DARKNESS, LIGHTED ROAD	FIXED OBJECT	ONE-WAY		STRAIGHT AND LEVEL	WET
35	JC415517		9/1/2019 1:00	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	RAIN	DARKNESS, LIGHTED ROAD	REAR END	DIVIDED - W/MEDIAN (NOT RAISED)		STRAIGHT AND LEVEL	WET
36	JC415625		9/1/2019 4:14	30	NO CONTROLS	NO CONTROLS	RAIN	DARKNESS, LIGHTED ROAD	PARKED MOTOR VEHICLE	NOT DIVIDED		STRAIGHT AND LEVEL	WET
37	JC412313		8/29/2019 16:10	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	RAIN	DAYLIGHT	SIDESWIPE SAME DIRECTION	ONE-WAY		STRAIGHT AND LEVEL	DRY
38	JC410534		8/28/2019 11:00	35	NO CONTROLS	NO CONTROLS	CLEAR	DAYLIGHT	REAR END	FOUR WAY		STRAIGHT AND LEVEL	DRY
39	JC415277		8/31/2019 20:40	30	TRAFFIC SIGNAL	UNKNOWN	RAIN	DARKNESS, LIGHTED ROAD	TURNING	NOT DIVIDED		STRAIGHT AND LEVEL	WET


Traffic_Crashes_-_Crashes_1 - Excel
Tahmid Alam
Traffic_Crashes_-_Crashes_1
Ready

M1 ROADWAY_SURFACE_COND

	M	N	O	P	Q	R	S	T	U	V
1	ROADWAY_SURFACE_COND	ROAD_DEFECT	REPORT_TYPE	CRASH_TYPE	INTERSECTION_RELATED_I	NOT_RIGHT_OF_WAY_I	HIT_AND_RUN_I	DAMAGE	DATE_POLICE_NOTIFIED	PRIM_CONTRIBUTORY_CAUSE
2	DRY	NO DEFECTS		NO INJURY / DRIVE AWAY				OVER $1,500	7/4/2019 23:05	FOLLOWING TOO CLOSELY
3	DRY	NO DEFECTS		NO INJURY / DRIVE AWAY				OVER $1,500	7/30/2019 10:25	FAILING TO YIELD RIGHT-OF-WAY
4	DRY	NO DEFECTS		NO INJURY / DRIVE AWAY				$501 - $1,500	8/10/2019 17:35	EQUIPMENT - VEHICLE CONDITION
5	DRY	NO DEFECTS		NO INJURY / DRIVE AWAY			Y	$501 - $1,500	8/16/2019 16:53	UNABLE TO DETERMINE
6	DRY	NO DEFECTS		NO INJURY / DRIVE AWAY			Y	$501 - $1,500	8/17/2019 18:30	UNABLE TO DETERMINE
7	DRY	UNKNOWN		NO INJURY / DRIVE AWAY	Y			OVER $1,500	8/28/2019 19:59	UNABLE TO DETERMINE
8	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY	Y			OVER $1,500	8/30/2019 14:25	FAILING TO REDUCE SPEED TO AVOID CRASH
9	DRY	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH				OVER $1,500	8/31/2019 4:35	PHYSICAL CONDITION OF DRIVER
10	UNKNOWN	UNKNOWN	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY				$501 - $1,500	8/30/2019 19:57	FOLLOWING TOO CLOSELY
11	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				OVER $1,500	8/31/2019 19:02	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
12	UNKNOWN	UNKNOWN	NOT ON SCENE (DESK REPORT)	INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	8/31/2019 16:49	FAILING TO REDUCE SPEED TO AVOID CRASH
13	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY	Y		Y	$500 OR LESS	8/29/2019 14:57	IMPROPER LANE USAGE
14	DRY	NO DEFECTS		INJURY AND / OR TOW DUE TO CRASH				OVER $1,500	8/30/2019 9:55	FAILING TO YIELD RIGHT-OF-WAY
15	WET	NO DEFECTS		INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	8/26/2019 8:37	IMPROPER OVERTAKING/PASSING
16	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				OVER $1,500	8/30/2019 18:19	UNABLE TO DETERMINE
17	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY			N	OVER $1,500	8/31/2019 11:15	UNABLE TO DETERMINE
18	UNKNOWN	UNKNOWN	NOT ON SCENE (DESK REPORT)	INJURY AND / OR TOW DUE TO CRASH				$501 - $1,500	8/31/2019 0:10	UNABLE TO DETERMINE
19	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				$500 OR LESS	8/31/2019 15:21	IMPROPER BACKING
20	WET	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	8/31/2019 23:15	NOT APPLICABLE
21	DRY	UNKNOWN	ON SCENE	NO INJURY / DRIVE AWAY			Y	$501 - $1,500	8/21/2019 19:50	UNABLE TO DETERMINE
22	UNKNOWN	UNKNOWN	ON SCENE	INJURY AND / OR TOW DUE TO CRASH				OVER $1,500	8/31/2019 0:46	IMPROPER OVERTAKING/PASSING
23	WET	UNKNOWN	ON SCENE	NO INJURY / DRIVE AWAY				$501 - $1,500	8/31/2019 18:34	IMPROPER BACKING
24	WET	UNKNOWN	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY	Y			$500 OR LESS	8/31/2019 22:50	FAILING TO YIELD RIGHT-OF-WAY
25	DRY	NO DEFECTS	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY		Y	N	$501 - $1,500	8/31/2019 14:35	UNABLE TO DETERMINE
26	DRY	UNKNOWN	ON SCENE	INJURY AND / OR TOW DUE TO CRASH	Y			OVER $1,500	8/25/2019 3:32	UNABLE TO DETERMINE
27	DRY	NO DEFECTS	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY		Y		$501 - $1,500	8/30/2019 11:35	IMPROPER BACKING
28	UNKNOWN	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY			Y	$500 OR LESS	8/31/2019 9:38	UNABLE TO DETERMINE
29	DRY	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	8/31/2019 10:35	UNABLE TO DETERMINE
30	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				OVER $1,500	8/31/2019 12:16	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
31	DRY	UNKNOWN		INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	8/31/2019 7:08	UNABLE TO DETERMINE
32	DRY	NO DEFECTS	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY	Y			$501 - $1,500	8/29/2019 18:20	IMPROPER BACKING
33	DRY	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH	Y			OVER $1,500	8/28/2019 11:20	FAILING TO YIELD RIGHT-OF-WAY
34	WET	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH				OVER $1,500	9/1/2019 0:47	DRIVING ON WRONG SIDE/WRONG WAY
35	WET	NO DEFECTS	NOT ON SCENE (DESK REPORT)	INJURY AND / OR TOW DUE TO CRASH			Y	OVER $1,500	9/1/2019 1:53	UNABLE TO DETERMINE
36	WET	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				$501 - $1,500	9/1/2019 4:14	WEATHER
37	DRY	NO DEFECTS	ON SCENE	NO INJURY / DRIVE AWAY				$501 - $1,500	8/29/2019 16:15	UNABLE TO DETERMINE
38	DRY	NO DEFECTS	ON SCENE	INJURY AND / OR TOW DUE TO CRASH	Y			OVER $1,500	8/28/2019 11:02	UNABLE TO DETERMINE
39	WET	UNKNOWN	NOT ON SCENE (DESK REPORT)	NO INJURY / DRIVE AWAY	Y			OVER $1,500	8/31/2019 21:13	UNABLE TO DETERMINE

Exploration of Collected data

- **Simple query:**
  - Counting the number of accidents occurred in a month, weather type, road type, time of the day and week
- **Getting the average on numeric data. E.g:** posted speed limit
- **Checking for null values for each variables**
- **Collection of unique values in categorical data**
  - **For weather:** Clear, Rain, Cloudy/Overcast, Snow etc.
  - **Lighting Condition:** Daylight, Darkness, Darkness-Lighted condition etc.
  - **First Crash:** Rear end, Parked, Sideswipe, Turning, etc.

2.3

Approach & Implementation

Data cleaning

- Very crucial step before any kind of analysis
- Removing the corrupted values
- Removing the rows with null or no values:
  - If missing value is less.

2.3

Approach & Implementation

Feature selection

- For **Kmeans algorithm** we have to find features
- After doing primary data exploration we selected **17 columns as features**
- Data can help **to create clusters and insight** for our purpose
- Some of the selected features are:
  1. Weather type
  2. Time of the day
  3. Day of the week
  4. Month
  5. Traffic way type etc.

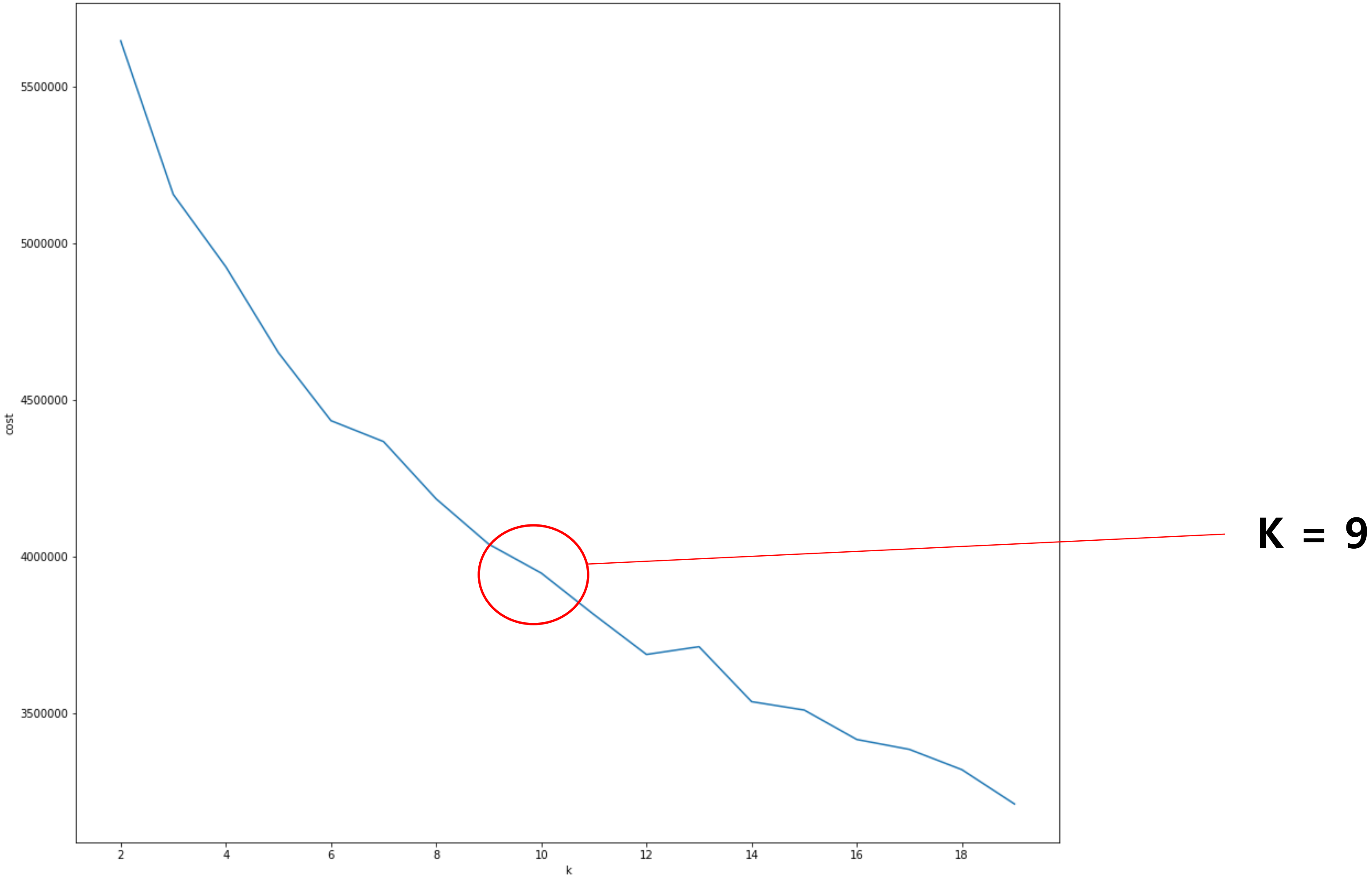
■ Elbow method

- Used to find out optimum number of K values for the data
- Uses wssse value as a metric
- Kmeans is run on from number of clusters from 2 to 20 or more
- Then using the wssse data a graph is generated
- After plotting this graph it looks like an elbow, hence the name

2.3

Approach & Implementation

- Elbow method

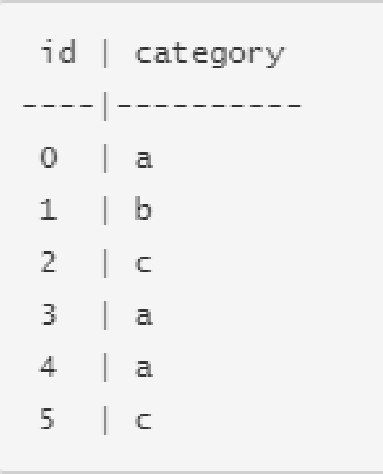


K = 9

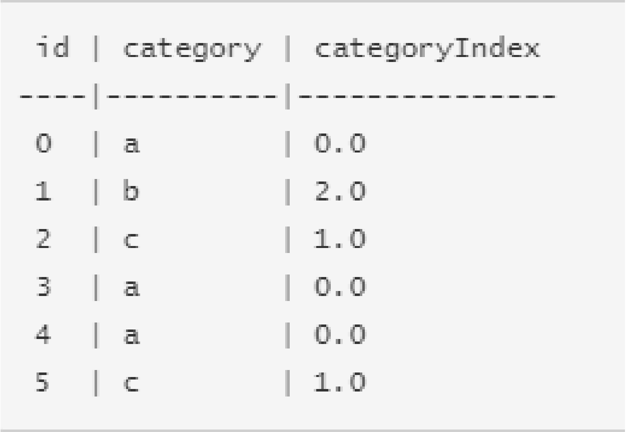
2.3
Approach & Implementation

Feature Indexing

- Spark Kmeans cannot work on categorical data.
- Have to convert the categorical data into numerical value.
- Using spark mlib library we can achieve that.
- "StringIndexer encodes a string column of labels to a column of label indices. The indices ar e in [0, numLabels), and four ordering options are supported."



```
 id | category
----|----------
 0  | a
 1  | b
 2  | c
 3  | a
 4  | a
 5  | c
```



```
 id | category | categoryIndex
----|----------|---------------
 0  | a        | 0.0
 1  | b        | 2.0
 2  | c        | 1.0
 3  | a        | 0.0
 4  | a        | 0.0
 5  | c        | 1.0
```

Source:https://spark.apache.org/docs/latest/ml-features

Feature Scaling

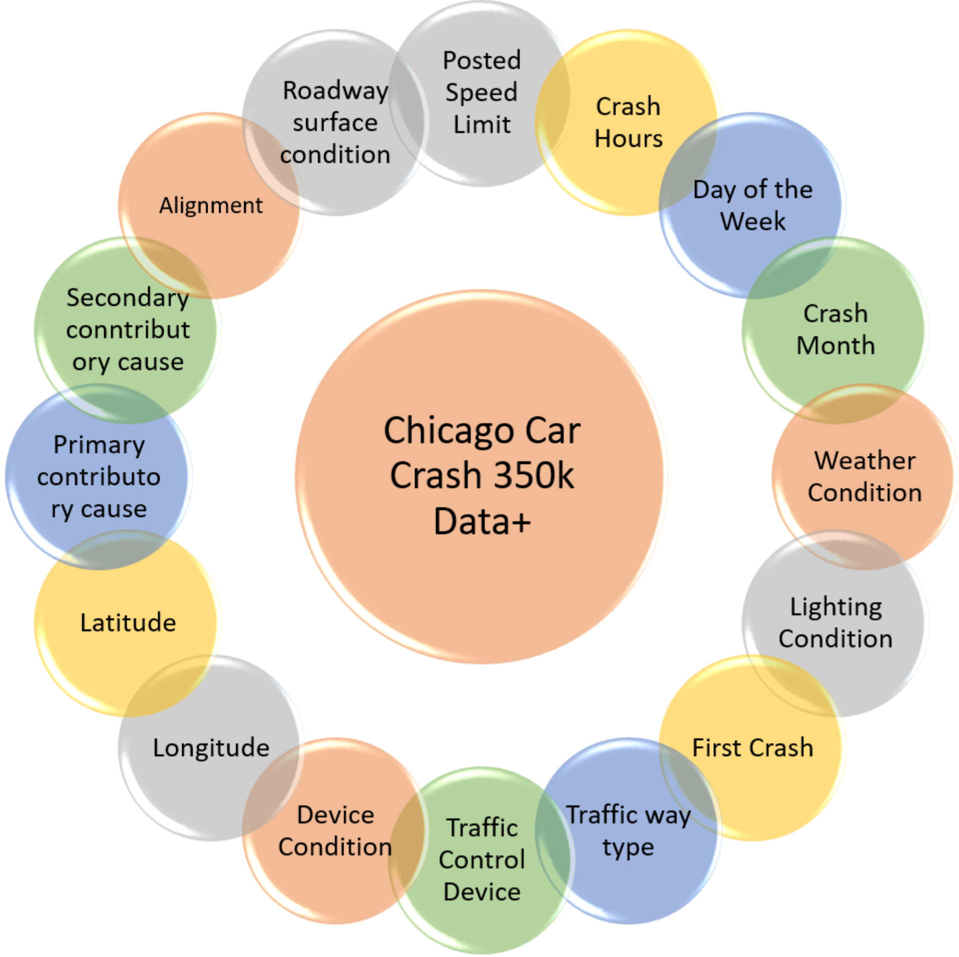
- Necessity of feature scaling
- Types of scaling (In spark mLib):
  - **Standard deviation**
  - **Mean**
- We used standard deviation method for our purpose

Source:https://spark.apache.org/docs/latest/ml-features

2.4
Approach & Implementation

K-means

- Unsupervised Machine Learning Algoritm
- Works with Eucledian distance
- Creates cluster based on closest centroid
- K = 9



3.1

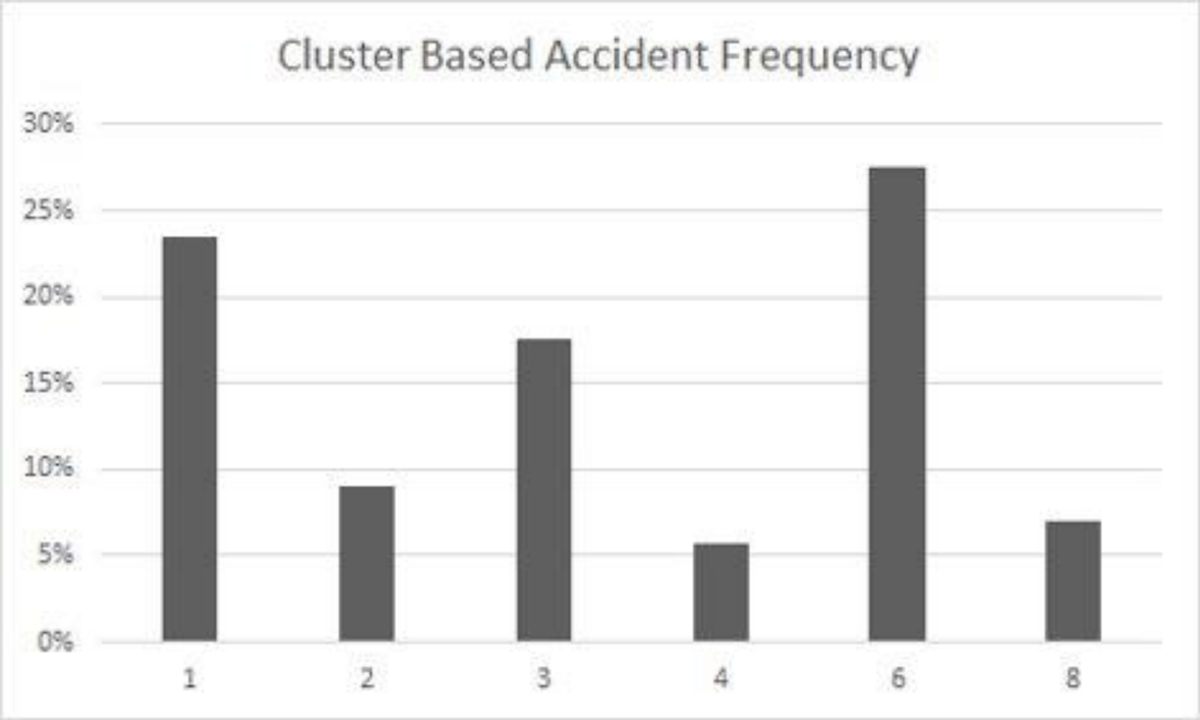
Insights and Recommendations

Cluster Identification

We divided our clusters into 5 regions:

- Low Region
- Low-Mid Region
- Mid Region
- High Region
- Very High Region

Cluster 1: High Region
Cluster 2 : Low-Mid Region
Cluster 3: Mid Region
Cluster 4 and 8: Low Region
Cluster 6: Very High Region



3.1
Insights and Recommendations

▪ Cluster 1: High Accident Frequency Region

Accident Count: 84,407

Cluster 1 accounts for 23% of the data set

Key Features:

- Weather Condition: Clear (89%)
- Lightening Condition: Daylight (71%) and Darkness-Lighted Road(20%)
- Crash Type: Rear End(36%), Sideswipe(17%) and Turning(24%)
- Traffic Way Type: Not Divided(50%) and Divided–without raised median(24%)
- Traffic Control Device: Traffic Signal(73%) and Stop sign/Flasher(19%)
- Device Condition: Functioning Properly(87%) and Functioning Improperly(1.5%)

3.1

Insights and Recommendations

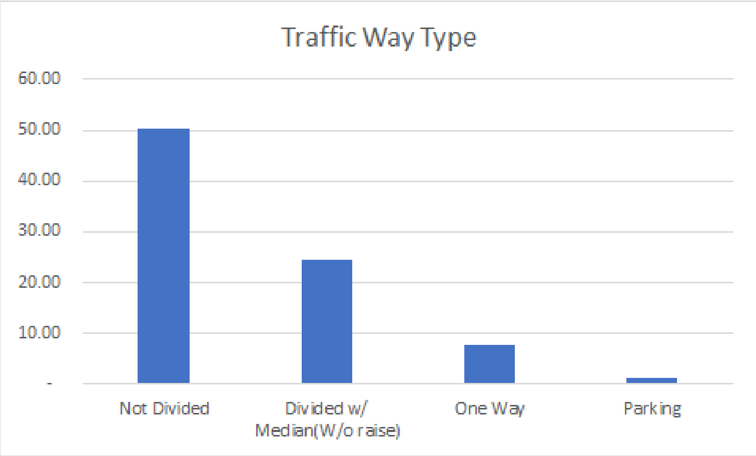
▪ Cluster 1: High Accident Frequency Region

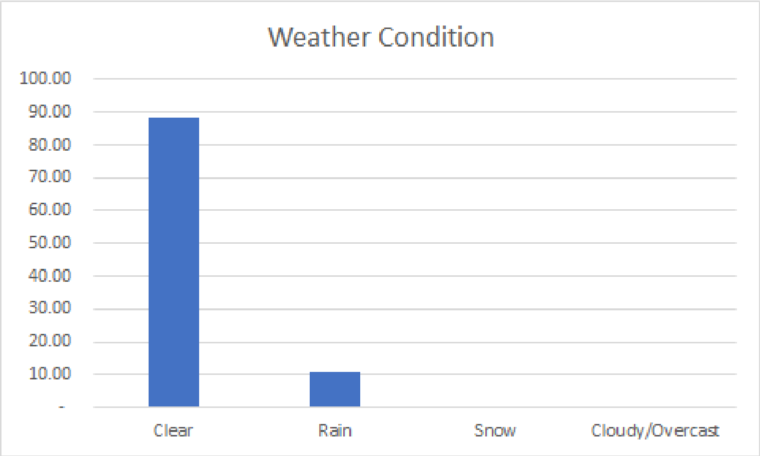
Insights:

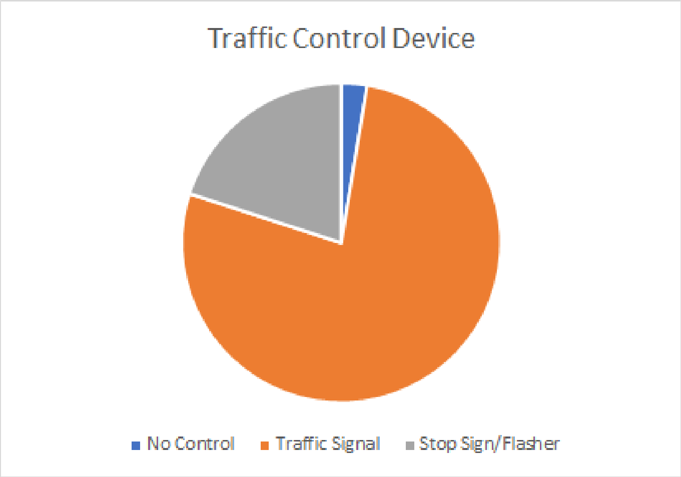
- Most Accidents happen on Non-Divided Roads
- Under Clear Weather and properly functional Traffic Control devices

Recommendations:

- Speed limit control in Non-Divided Roads
- Caution Signs







Cluster 2: Low-Mid Accident Frequency Region

Accident Count: 32,625

Cluster 2 accounts for only 9% of the data set

Key Features:

- Weather Condition: Snow(35%), Cloudy/Overcast(29%)
- Lightening Condition: Daylight (54%) and Darkness-Lighted Road(20%)
- Crash Type: Parked(33%) and Rear End(23%)
- Traffic Way Type: Not Divided(44%), Divided–w/o raised median(17%) and Oneway(18%)
- Traffic Control Device: No Control(66%) and Traffic Signal(23%) and Stop sign/Flasher(10%)
- Device Condition: No Control(66%) and Functioning Properly(30%)

3.2
Insights and Recommendations

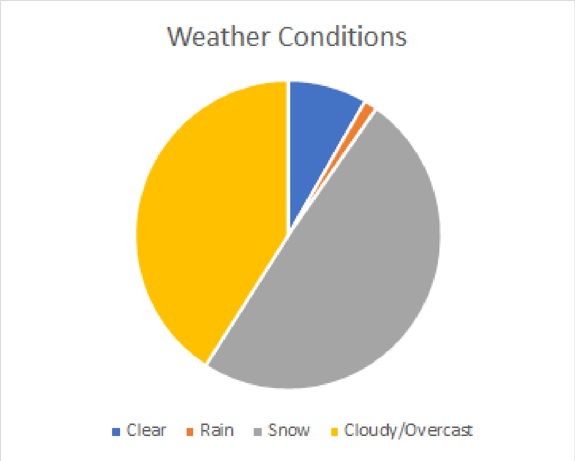
Cluster 2: Low-Mid Accident Frequency Region

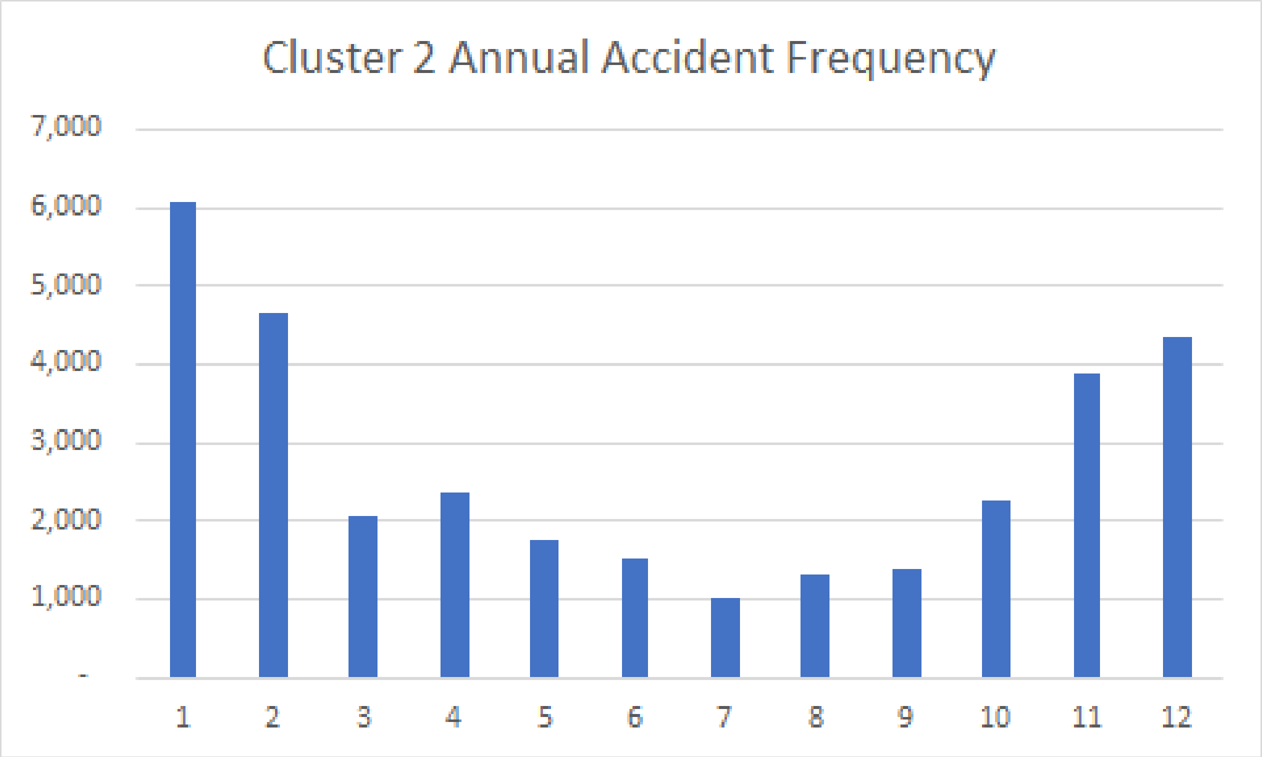
Insights:

- Most Accidents Occurred in Winters
- Non-Divided and One-Way Roads
- No Control Device Present

Recommendations:

- Have traffic control devices/stop signs on these roads





▪ Cluster 3: Mid Accident Frequency Region

Accident Count: 63,419

Cluster 3 accounts for 18% of the data set

Key Features:

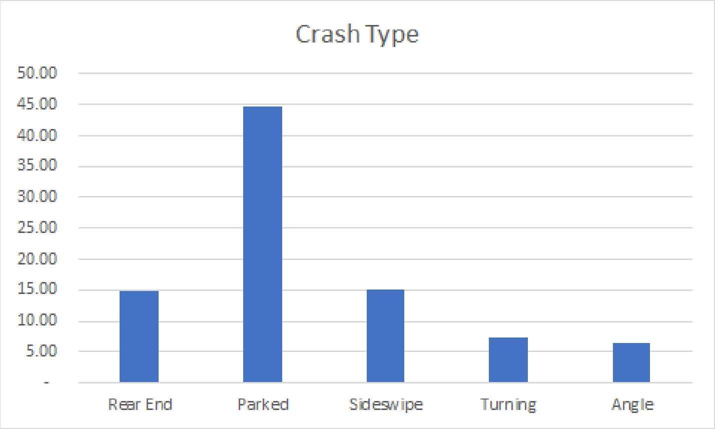
- Weather Condition: Clear(90%) and Rain(9%)
- Lightening Condition: Daylight (67%) and Darkness-Lighted Road(21%)
- Crash Type: Parked(45%), Rear End(15%) and Side Swipe(15%)
- Traffic Way Type: Not Divided(46%), Oneway(25%) and Parking(4%)
- Traffic Control Device: No Control(99.4%)
- Device Condition: No Control(99.4%)

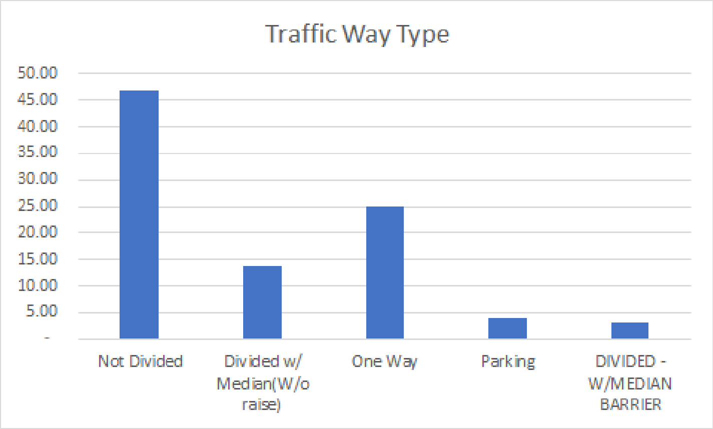
3.3
Insights and Recommendations

▪ Cluster 3: Mid Accident Frequency Region

Insights:
- Most Accidents happen on Non-Divided or One-Way Roads
- Under Clear Weather but with no Traffic Control device
- Major accident type is Parked Motor Vehicle but that does not happen in Parking area

Recommendations:
- Traffic Control Devices
- Caution Signs for Non-Divided and One-Way roads that have roadside parking





3.4
Insights and Recommendations

■ Cluster 4: Low Accident Frequency Region

Insights:

- Most Accidents happen on Non-Divided
- Under Clear Weather, with and without Traffic Control device

Recommendations:

- More Traffic Control Devices on Non-Divided roads to prevent these kinds of accidents

3.5
Insights and Recommendations

▪ Cluster 6: Very High Accident Frequency Region

Accident Count: 99,012 (Cluster with Maximum Accidents)

Cluster 6 accounts for 28% of the data set

Key Features:

- Weather Condition: Clear(91%) and Rain(9%)
- Lightening Condition: Daylight (74%) and Darkness-Lighted Road(16%)
- Crash Type: Parked(26%), Side Swipe(21%) and Rear End(23%)
- Traffic Way Type: Not Divided(43%), Divided–w/o raised median(17%) and Parking(18%)
- Traffic Control Device: No Control(97%)
- Device Condition: No Control(98%)

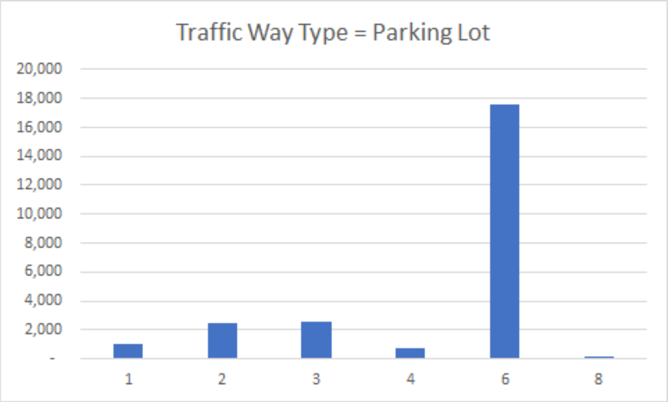
Cluster 6: Very High Accident Frequency Region

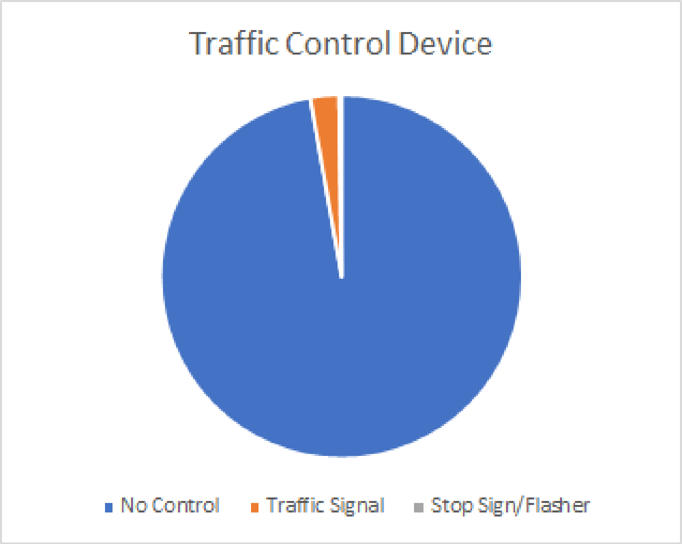
Insights:

- Most Parking Lot Accidents in this cluster
- Under Clear Weather and without Traffic Control device
- 13,586 out of 17,595 (77%) accidents happened in Parking lot during Daytime

Recommendations:

- People need to be extra cautious in Parking Lots
- Sign and signal inside Parking Lots





▪ Cluster 4: Low Accident Frequency Region

Accident Count: 20,673

Cluster 4 accounts for 6% of the data set

Key Features:

- Weather Condition: Clear(83%) and Rain(14%)
- Lightening Condition: Daylight (48%) and Darkness-Lighted Road(37%)
- Crash Type: Parked(17.5%), Angle(17.5%) and Rear End(14%)
- Traffic Way Type: Not Divided(42%) and Divided–w/o raised median(20%)
- Traffic Control Device: No Control(51%), Traffic Signal(33%) and Stop Sign(13%)
- Device Condition: No Control(53%) and Functioning Properly(44%)

3.6

Insights and Recommendations

▪ Cluster 8: Low Accident Frequency Region

Accident Count: 25,340

Cluster 8 accounts for just 7% of the data set

Key Features:

- Weather Condition: Clear(86%) and Rain(11%)
- Lightening Condition: Daylight (71%) and Darkness-Lighted Road(20%)
- Crash Type: Turning(15%), Side Swipe(15%) and Rear End(33%)
- Traffic Way Type: Not Divided(50%), Divided–w/o raised median(22%) and One-Way(14%)
- Traffic Control Device: Traffic Signal(45%) and Stop Sign/Flasher(35%)
- Device Condition: Functioning Properly(74%)

Cluster 8: Low Accident Frequency Region

Insights:

- Control Devices were functioning properly
- Turning, Side Swipe and Rear-end are major accident types
- Clearly shows human factor as major cause of accident

Recommendations:

- People need to be extra cautious when taking a turn or overtaking

4.1
Summary and Conclusion

Summary

- In the **highest** accident frequency region: People need to be extra cautious in **Parking Lots**
- In **high accident** frequency region: **Speed limit control in Non-Divided Roads**, Caution Signs
- In **low-mid accident** frequency region: Have **traffic control** devices/**stop signs**
- In **mid accident** frequency region: **Caution Signs for Non-Divided** and **One-Way roads** that have roadside parking
- In **low accident** frequency region: **More Traffic Control Devices** on Non-Divided roads

4.2

Summary and Conclusion

■ Limitations & Difficulties

- Big data -> take lots of time for data sorting
- Hardware Limitation: running software with a large data in virtual box would often crash
- Jupyter kernal frequently crashed while running the program
- Missing data manipulation library
- Learning curve

4.3
Summary and Conclusion

▪ Future work

- With this analyzed data, the project can be improved as **a prediction of car crash**
- The project can add advanced **machine Learning algorithms**
- More functionalities
- Suggest the outcomes or summary to Stakeholders (Police department, hospital and insurance company etc)

CS590 Big Data and Cloud Computing

Q & A

Min Namgung & Shehryar Shahid & Tahmid Alam