ProjectProgress(1)

August 10, 2023

1 IST652 Project Deliverable 2

1.1 Phase 2: Project Progress

In this step you should have a road map of the steps you will be taking to complete your analysis. In addition, at this stage you should also complete the following - fine tune your research questions. - upload your dataset into jupyterhub and conduct some preliminary cleaning and transformation. - provide coding activities conducted so far. - have a better sense of team members responsibilities. - set a schedule to meet

1.1.1 Step 1: What is Your Idea and Research Questions, Again?

Please reiterate your project idea below (you can copy it from the project proposal if there were no changes).

The project's focus on analyzing authoritative data from the NYPD and addressing critical questions related to traffic safety in NYC aligns with the overall goal of improving road safety and decision-making for stakeholders and policymakers and our goal remains the same as we presented in Project Proposal with some updated research questions.

1. Which vehicle types are involved in the highest number of collisions, and do certain types have a higher likelihood of being in severe accidents compared to others?
2. What are the most common contributing factors to collisions, and do these factors vary based on vehicle characteristics or driver demographics?
3. Is there any correlation between the age of vehicles and the severity of collisions they are involved in?
4. Which specific area or neighborhood in New York State experienced the highest number of vehicle collisions due to accidents?
5. How does driver gender impact collision rates and severity? Are there notable differences between male and female drivers in terms of the types of collisions they are involved in?
6. What is the maximum number of vehicles involved in a single collision, and how many collisions had this maximum number of vehicle intersections?
7. During which time period did the majority of accidents occur in New York State?
8. Which part of New York state had seen most vehicle collisions due to accidents.

9. What are the common situations or events that occurred immediately before collisions? Are there any patterns in pre-crash events that lead to different levels of property damage?
10. What is the relationship between driver's license status and the frequency of accidents? Are certain driver's license statuses associated with a higher incidence of accidents?
11. Which types of vehicle damages are most prevalent in collisions, and are there correlations with specific contributing factors?

1.1.2 Step 2: Problem Analysis - Roadmap

What are the preliminary major steps you will be completing? Include the research question and steps taken to answer that question? Are there any unique functions you will be incorporating which we have not covered in the classroom? Describe below.

The preliminary major steps taken in the code provided include data preprocessing, data cleaning, and initial exploratory data analysis.

It demonstrates an initial exploration of the dataset to address some of the research questions. However, there are further steps required to fully analyze and answer all the research questions. These additional steps may involve more advanced statistical analysis, data visualization, and potentially may be machine learning techniques to uncover insights and patterns related to the factors influencing motor vehicle collisions in NYC.

The code utilizes various pandas functions and there are some novel functions incorporated which are cut functions to create bins for time intervals and also created a list for multiple damages which consisted of 3 variables into single column.

1.1.3 Step 3: Preliminary Code

Include coding that has been completed at this preliminary stage.

[1]:
```
#Importing pyhton libraries
%matplotlib inline

import pandas as pd
import numpy as np
import requests
from io import StringIO
from io import BytesIO
from zipfile import ZipFile
import re
#Add additional libraries below this line
import seaborn as sns
from matplotlib import pyplot as plt
np.set_printoptions(precision=4)
pd.options.display.max_rows = 25
```

[2]:
```
filepath="~/datasets/ist652/Summer2023/Project_Data_.csv"

data= pd.read_csv(filepath)
```

```
data.head()
```

[2]:

	UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE	VEHICLE_MODEL	VEHICLE_YEAR	…	PRE_CRASH	POINT_OF_IMPACT	VEHICLE_DAMAGE	VEHICLE_DAMAGE_1	VEHICLE_DAMAGE_2	VEHICLE_DAMAGE_3	PUBLIC_PROPERTY_DAMAGE	PUBLIC_PROPERTY_DAMAGE_TYPE	CONTRIBUTING_FACTOR_1	CONTRIBUTING_FACTOR_2
0	19140702	4213082	09/23/2019	8:15	0553ab4d-9500-4cba-8d98-f4d7f89d5856	NY	Station Wagon/Sport Utility Vehicle	TOYT -CAR/SUV	NaN	2002.0	…	Going Straight Ahead	Left Front Bumper	Left Front Quarter Panel	NaN	NaN	NaN	N	NaN	Driver Inattention/Distraction	Unspecified
1	17044639	3434155	05/02/16	17:35	219456	NY	4 dr sedan	MERZ -CAR/SUV	NaN	2015.0	…	Merging	Right Front Bumper	Right Front Bumper	Right Front Quarter Panel	NaN	NaN	N	NaN	Driver Inattention/Distraction	Unsafe Lane Changing
2	19138701	4229067	10/24/2019	13:15	c53b43d9-419a-4ab1-9361-3f2979078d89	NY	Bus	FRHT-TRUCK/BUS	NaN	2006.0	…	Parked	Left Front Quarter Panel	Left Front Quarter Panel	NaN	NaN	NaN	N	NaN		
3	17303317	3503027	08/18/2016	12:39	672828	NY	Station Wagon/Sport Utility Vehicle	FORD -CAR/SUV	NaN	2005.0	…	Going Straight Ahead	Center Front End	Center Front End	No Damage	No Damage	No Damage	N	NaN		
4	17285715	3487936	07/22/2016	15:40	554272	NY	Convertible	VOLK -CAR/SUV	NaN	2013.0	…	Stopped in Traffic	Right Rear Bumper	Right Rear Bumper	Center Back End	Left Rear Bumper	NaN	N	NaN		

```
2                     Unspecified           Unspecified
3  Driver Inattention/Distraction           Unspecified
4                     Unspecified           Unspecified

[5 rows x 25 columns]
```

[3]:
```
data.shape
```

[3]:
```
(154271, 25)
```

[4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 154271 entries, 0 to 154270
Data columns (total 25 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   UNIQUE_ID                      154271 non-null  int64
 1   COLLISION_ID                   154271 non-null  int64
 2   CRASH_DATE                     154271 non-null  object
 3   CRASH_TIME                     154271 non-null  object
 4   VEHICLE_ID                     154271 non-null  object
 5   STATE_REGISTRATION             153470 non-null  object
 6   VEHICLE_TYPE                   154267 non-null  object
 7   VEHICLE_MAKE                   151260 non-null  object
 8   VEHICLE_MODEL                  22091 non-null   object
 9   VEHICLE_YEAR                   152086 non-null  float64
 10  TRAVEL_DIRECTION               154255 non-null  object
 11  VEHICLE_OCCUPANTS              154267 non-null  float64
 12  DRIVER_SEX                     154271 non-null  object
 13  DRIVER_LICENSE_STATUS          154271 non-null  object
 14  DRIVER_LICENSE_JURISDICTION    152663 non-null  object
 15  PRE_CRASH                      153575 non-null  object
 16  POINT_OF_IMPACT                154261 non-null  object
 17  VEHICLE_DAMAGE                 154253 non-null  object
 18  VEHICLE_DAMAGE_1               91401 non-null   object
 19  VEHICLE_DAMAGE_2               64621 non-null   object
 20  VEHICLE_DAMAGE_3               47494 non-null   object
 21  PUBLIC_PROPERTY_DAMAGE         154271 non-null  object
 22  PUBLIC_PROPERTY_DAMAGE_TYPE    0 non-null       float64
 23  CONTRIBUTING_FACTOR_1          154256 non-null  object
 24  CONTRIBUTING_FACTOR_2          152158 non-null  object
dtypes: float64(3), int64(2), object(20)
memory usage: 29.4+ MB
```

2 Cleaning the Dataset 1

[5]:
```
data.head() # Previewing the dataset 1.
```

[5]:
```
   UNIQUE_ID  COLLISION_ID  CRASH_DATE CRASH_TIME  \
0   19140702       4213082  09/23/2019       8:15
1   17044639       3434155    05/02/16      17:35
2   19138701       4229067  10/24/2019      13:15
3   17303317       3503027  08/18/2016      12:39
4   17285715       3487936  07/22/2016      15:40

                             VEHICLE_ID STATE_REGISTRATION  \
0  0553ab4d-9500-4cba-8d98-f4d7f89d5856                 NY
1                                219456                 NY
2  c53b43d9-419a-4ab1-9361-3f2979078d89                 NY
3                                672828                 NY
4                                554272                 NY

                         VEHICLE_TYPE    VEHICLE_MAKE VEHICLE_MODEL  \
0  Station Wagon/Sport Utility Vehicle   TOYT -CAR/SUV           NaN
1                          4 dr sedan   MERZ -CAR/SUV           NaN
2                                 Bus  FRHT-TRUCK/BUS           NaN
3  Station Wagon/Sport Utility Vehicle   FORD -CAR/SUV           NaN
4                          Convertible   VOLK -CAR/SUV           NaN

   VEHICLE_YEAR  …              PRE_CRASH           POINT_OF_IMPACT  \
0        2002.0  …  Going Straight Ahead         Left Front Bumper
1        2015.0  …                Merging        Right Front Bumper
2        2006.0  …                 Parked  Left Front Quarter Panel
3        2005.0  …  Going Straight Ahead          Center Front End
4        2013.0  …    Stopped in Traffic         Right Rear Bumper

            VEHICLE_DAMAGE           VEHICLE_DAMAGE_1  VEHICLE_DAMAGE_2  \
0  Left Front Quarter Panel                        NaN               NaN
1       Right Front Bumper  Right Front Quarter Panel               NaN
2  Left Front Quarter Panel                        NaN               NaN
3         Center Front End                  No Damage         No Damage
4        Right Rear Bumper            Center Back End  Left Rear Bumper

  VEHICLE_DAMAGE_3 PUBLIC_PROPERTY_DAMAGE PUBLIC_PROPERTY_DAMAGE_TYPE  \
0              NaN                      N                         NaN
1              NaN                      N                         NaN
2              NaN                      N                         NaN
3        No Damage                      N                         NaN
4              NaN                      N                         NaN

           CONTRIBUTING_FACTOR_1 CONTRIBUTING_FACTOR_2
```

```
0  Driver Inattention/Distraction           Unspecified
1  Driver Inattention/Distraction  Unsafe Lane Changing
2                     Unspecified           Unspecified
3  Driver Inattention/Distraction           Unspecified
4                     Unspecified           Unspecified

[5 rows x 25 columns]
```

2.0.1 converting date and time columns to appropriate formats

[6]:
```
data['CRASH_DATE'] = pd.to_datetime(data['CRASH_DATE'])
date_format = '%Y-%m-%d'
data.sort_values(by=['CRASH_DATE'], inplace=True)# Sorting the data based on␣
 ↪the year in which the car was registered.
data=data.reset_index(drop=True)

data['CRASH_TIME'] = pd.to_datetime(data['CRASH_TIME'], format='%H:%M')
data['CRASH_TIME'] = data['CRASH_TIME'].dt.strftime('%H:%M')
data.head()
```

[6]:
```
   UNIQUE_ID  COLLISION_ID CRASH_DATE CRASH_TIME VEHICLE_ID  \
0   17307404       3493857 2012-07-31      13:30     591401
1   17295384       3518768 2012-09-12      12:20     780983
2   17218669       3576809 2012-12-09      12:03    1178874
3   17143961       3421148 2013-04-05      09:40     453654
4   17272443       3467155 2013-06-23      08:30     437750

  STATE_REGISTRATION                         VEHICLE_TYPE   VEHICLE_MAKE  \
0                 NY                            Box Truck  GMC-TRUCK/BUS
1                 NY                                Sedan  HOND -CAR/SUV
2                 NJ                                 Taxi  CADI -CAR/SUV
3                 NY  Station Wagon/Sport Utility Vehicle  FORD -CAR/SUV
4                 NY                              4 dr sedan  TOYT -CAR/SUV

  VEHICLE_MODEL  VEHICLE_YEAR  …                PRE_CRASH  \
0           NaN        1997.0  …            Changing Lanes
1           NaN        2000.0  …          Making Left Turn
2           NaN        2015.0  …  Entering Parked Position
3     FORD ECP         2015.0  …                  Backing
4           NaN        2012.0  …      Going Straight Ahead

            POINT_OF_IMPACT            VEHICLE_DAMAGE  VEHICLE_DAMAGE_1  \
0          Right Side Doors          Right Side Doors         No Damage
1         Left Front Bumper         Left Front Bumper  Center Front End
2  Right Rear Quarter Panel  Right Rear Quarter Panel         No Damage
3  Left Front Quarter Panel  Left Front Quarter Panel               NaN
4         Right Rear Bumper         Right Rear Bumper   Center Back End
```

```
           VEHICLE_DAMAGE_2 VEHICLE_DAMAGE_3 PUBLIC_PROPERTY_DAMAGE  \
0                No Damage        No Damage                       N
1  Left Front Quarter Panel              NaN                       N
2                No Damage        No Damage                       N
3                      NaN              NaN                       N
4          Left Rear Bumper        No Damage                       N

  PUBLIC_PROPERTY_DAMAGE_TYPE           CONTRIBUTING_FACTOR_1  \
0                         NaN  Driver Inattention/Distraction
1                         NaN             Driver Inexperience
2                         NaN             Passing Too Closely
3                         NaN                     Unspecified
4                         NaN                     Unspecified

  CONTRIBUTING_FACTOR_2
0           Unspecified
1    Turning Improperly
2           Unspecified
3           Unspecified
4           Unspecified

[5 rows x 25 columns]
```

2.0.2 Convert 'VEHICLE_YEAR' data type to int format

[7]:
```
data['VEHICLE_YEAR'].fillna(0, inplace=True)
data['VEHICLE_YEAR'] = data['VEHICLE_YEAR'].astype(int)
data.head()
```

[7]:
```
    UNIQUE_ID  COLLISION_ID CRASH_DATE CRASH_TIME VEHICLE_ID  \
0   17307404       3493857 2012-07-31      13:30     591401
1   17295384       3518768 2012-09-12      12:20     780983
2   17218669       3576809 2012-12-09      12:03    1178874
3   17143961       3421148 2013-04-05      09:40     453654
4   17272443       3467155 2013-06-23      08:30     437750

  STATE_REGISTRATION                         VEHICLE_TYPE   VEHICLE_MAKE  \
0                 NY                            Box Truck  GMC-TRUCK/BUS
1                 NY                                Sedan  HOND -CAR/SUV
2                 NJ                                 Taxi  CADI -CAR/SUV
3                 NY  Station Wagon/Sport Utility Vehicle  FORD -CAR/SUV
4                 NY                            4 dr sedan  TOYT -CAR/SUV

  VEHICLE_MODEL  VEHICLE_YEAR  …                PRE_CRASH  \
0           NaN          1997  …          Changing Lanes
1           NaN          2000  …        Making Left Turn
```

```
2           NaN          2015  …  Entering Parked Position
3      FORD ECP          2015  …                  Backing
4           NaN          2012  …      Going Straight Ahead

            POINT_OF_IMPACT            VEHICLE_DAMAGE  VEHICLE_DAMAGE_1  \
0          Right Side Doors          Right Side Doors         No Damage
1         Left Front Bumper         Left Front Bumper  Center Front End
2  Right Rear Quarter Panel  Right Rear Quarter Panel         No Damage
3  Left Front Quarter Panel  Left Front Quarter Panel               NaN
4         Right Rear Bumper         Right Rear Bumper   Center Back End

           VEHICLE_DAMAGE_2 VEHICLE_DAMAGE_3 PUBLIC_PROPERTY_DAMAGE  \
0                No Damage        No Damage                      N
1  Left Front Quarter Panel              NaN                      N
2                No Damage        No Damage                      N
3                      NaN              NaN                      N
4          Left Rear Bumper        No Damage                      N

  PUBLIC_PROPERTY_DAMAGE_TYPE           CONTRIBUTING_FACTOR_1  \
0                         NaN  Driver Inattention/Distraction
1                         NaN             Driver Inexperience
2                         NaN             Passing Too Closely
3                         NaN                     Unspecified
4                         NaN                     Unspecified

  CONTRIBUTING_FACTOR_2
0           Unspecified
1    Turning Improperly
2           Unspecified
3           Unspecified
4           Unspecified

[5 rows x 25 columns]
```

2.0.3 Convert 'VEHICLE_YEAR' data type to int format

[8]:
```
data['VEHICLE_OCCUPANTS'].fillna(0, inplace=True)
data['VEHICLE_OCCUPANTS'] = data['VEHICLE_OCCUPANTS'].astype(int)
data.head()
```

[8]:
```
   UNIQUE_ID  COLLISION_ID CRASH_DATE CRASH_TIME VEHICLE_ID  \
0   17307404       3493857 2012-07-31      13:30     591401
1   17295384       3518768 2012-09-12      12:20     780983
2   17218669       3576809 2012-12-09      12:03    1178874
3   17143961       3421148 2013-04-05      09:40     453654
4   17272443       3467155 2013-06-23      08:30     437750
```

```
  STATE_REGISTRATION                          VEHICLE_TYPE   VEHICLE_MAKE  \
0                 NY                             Box Truck  GMC-TRUCK/BUS
1                 NY                                 Sedan  HOND -CAR/SUV
2                 NJ                                  Taxi  CADI -CAR/SUV
3                 NY  Station Wagon/Sport Utility Vehicle  FORD -CAR/SUV
4                 NY                              4 dr sedan  TOYT -CAR/SUV

  VEHICLE_MODEL  VEHICLE_YEAR  …                 PRE_CRASH  \
0           NaN          1997  …             Changing Lanes
1           NaN          2000  …           Making Left Turn
2           NaN          2015  …  Entering Parked Position
3     FORD ECP           2015  …                   Backing
4           NaN          2012  …      Going Straight Ahead

           POINT_OF_IMPACT            VEHICLE_DAMAGE  VEHICLE_DAMAGE_1  \
0          Right Side Doors          Right Side Doors         No Damage
1         Left Front Bumper         Left Front Bumper  Center Front End
2  Right Rear Quarter Panel  Right Rear Quarter Panel         No Damage
3  Left Front Quarter Panel  Left Front Quarter Panel               NaN
4         Right Rear Bumper         Right Rear Bumper   Center Back End

          VEHICLE_DAMAGE_2 VEHICLE_DAMAGE_3 PUBLIC_PROPERTY_DAMAGE  \
0                No Damage        No Damage                       N
1  Left Front Quarter Panel              NaN                       N
2                No Damage        No Damage                       N
3                      NaN              NaN                       N
4          Left Rear Bumper        No Damage                       N

  PUBLIC_PROPERTY_DAMAGE_TYPE           CONTRIBUTING_FACTOR_1  \
0                         NaN  Driver Inattention/Distraction
1                         NaN             Driver Inexperience
2                         NaN             Passing Too Closely
3                         NaN                     Unspecified
4                         NaN                     Unspecified

  CONTRIBUTING_FACTOR_2
0           Unspecified
1    Turning Improperly
2           Unspecified
3           Unspecified
4           Unspecified

[5 rows x 25 columns]
```

2.0.4 Convert all vehicle damages into single column as comma-seperated list

[9]:
```
data['VEHICLE_DAMAGE']=data['VEHICLE_DAMAGE'].replace({'No Damage':''})
data['VEHICLE_DAMAGE_1']=data['VEHICLE_DAMAGE_1'].replace({'No Damage':''})
data['VEHICLE_DAMAGE_2']=data['VEHICLE_DAMAGE_2'].replace({'No Damage':''})
data['VEHICLE_DAMAGE_3']=data['VEHICLE_DAMAGE_3'].replace({'No Damage':''})
data['VEHICLE_DAMAGE']= data['VEHICLE_DAMAGE'].fillna('')
data['VEHICLE_DAMAGE_1']= data['VEHICLE_DAMAGE_1'].fillna('')
data['VEHICLE_DAMAGE_2']= data['VEHICLE_DAMAGE_2'].fillna('')
data['VEHICLE_DAMAGE_3']= data['VEHICLE_DAMAGE_3'].fillna('')
data['VEHICLE_DAMAGES'] =␣
 ↪data["VEHICLE_DAMAGE"]+','+data['VEHICLE_DAMAGE_1']+','+data['VEHICLE_DAMAGE_2']+','+data['


def remove_repeated_commas(text):
    return re.sub(r'^,+|,+$|,(?=,)', '', text)
data['VEHICLE_DAMAGES'] = data['VEHICLE_DAMAGES'].apply(remove_repeated_commas)
data.head()
```

[9]:

	UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE	VEHICLE_MODEL	VEHICLE_YEAR	…	POINT_OF_IMPACT	VEHICLE_DAMAGE	VEHICLE_DAMAGE_1	VEHICLE_DAMAGE_2
0	17307404	3493857	2012-07-31	13:30	591401	NY	Box Truck	GMC-TRUCK/BUS	NaN	1997	…	Right Side Doors	Right Side Doors		
1	17295384	3518768	2012-09-12	12:20	780983	NY	Sedan	HOND -CAR/SUV	NaN	2000	…	Left Front Bumper	Left Front Bumper	Center Front End	Left Front Quarter Panel
2	17218669	3576809	2012-12-09	12:03	1178874	NJ	Taxi	CADI -CAR/SUV	NaN	2015	…	Right Rear Quarter Panel	Right Rear Quarter Panel		
3	17143961	3421148	2013-04-05	09:40	453654	NY	Station Wagon/Sport Utility Vehicle	FORD -CAR/SUV	FORD ECP	2015	…	Left Front Quarter Panel	Left Front Quarter Panel		
4	17272443	3467155	2013-06-23	08:30	437750	NY	4 dr sedan	TOYT -CAR/SUV	NaN	2012	…	Right Rear Bumper	Right Rear Bumper	Center Back End	Left Rear Bumper

```
  VEHICLE_DAMAGE_3 PUBLIC_PROPERTY_DAMAGE PUBLIC_PROPERTY_DAMAGE_TYPE  \
0                                       N                          NaN
1                                       N                          NaN
2                                       N                          NaN
3                                       N                          NaN
4                                       N                          NaN

            CONTRIBUTING_FACTOR_1 CONTRIBUTING_FACTOR_2  \
0  Driver Inattention/Distraction           Unspecified
1              Driver Inexperience    Turning Improperly
2              Passing Too Closely           Unspecified
3                     Unspecified           Unspecified
4                     Unspecified           Unspecified

                                    VEHICLE_DAMAGES
0                                   Right Side Doors
1  Left Front Bumper,Center Front End,Left Front …
2                           Right Rear Quarter Panel
3                           Left Front Quarter Panel
4  Right Rear Bumper,Center Back End,Left Rear Bu…

[5 rows x 26 columns]
```

2.0.5 Drop the 4 vehicle damages columns after combining into single column

[10]:
```
data = data.
↪drop(['VEHICLE_DAMAGE','VEHICLE_DAMAGE_1','VEHICLE_DAMAGE_2','VEHICLE_DAMAGE_3'],␣
↪axis = 1)
data.head()
```

[10]:
```
   UNIQUE_ID  COLLISION_ID CRASH_DATE CRASH_TIME VEHICLE_ID  \
0   17307404       3493857 2012-07-31      13:30     591401
1   17295384       3518768 2012-09-12      12:20     780983
2   17218669       3576809 2012-12-09      12:03    1178874
3   17143961       3421148 2013-04-05      09:40     453654
4   17272443       3467155 2013-06-23      08:30     437750

  STATE_REGISTRATION                       VEHICLE_TYPE   VEHICLE_MAKE  \
0                 NY                          Box Truck  GMC-TRUCK/BUS
1                 NY                              Sedan  HOND -CAR/SUV
2                 NJ                               Taxi  CADI -CAR/SUV
3                 NY  Station Wagon/Sport Utility Vehicle  FORD -CAR/SUV
4                 NY                          4 dr sedan  TOYT -CAR/SUV

  VEHICLE_MODEL  VEHICLE_YEAR  … DRIVER_SEX  DRIVER_LICENSE_STATUS  \
0           NaN          1997  …          M                Licensed
1           NaN          2000  …          M                Licensed
```

```
2           NaN          2015  …          M               Licensed
3      FORD ECP          2015  …          M               Licensed
4           NaN          2012  …          F               Licensed

  DRIVER_LICENSE_JURISDICTION                PRE_CRASH  \
0                          NY            Changing Lanes
1                          NY          Making Left Turn
2                          NJ  Entering Parked Position
3                          NY                  Backing
4                          NY      Going Straight Ahead

           POINT_OF_IMPACT PUBLIC_PROPERTY_DAMAGE  \
0          Right Side Doors                      N
1         Left Front Bumper                      N
2  Right Rear Quarter Panel                      N
3  Left Front Quarter Panel                      N
4         Right Rear Bumper                      N

  PUBLIC_PROPERTY_DAMAGE_TYPE           CONTRIBUTING_FACTOR_1  \
0                         NaN  Driver Inattention/Distraction
1                         NaN            Driver Inexperience
2                         NaN            Passing Too Closely
3                         NaN                    Unspecified
4                         NaN                    Unspecified

   CONTRIBUTING_FACTOR_2                                    VEHICLE_DAMAGES
0           Unspecified                                      Right Side Doors
1     Turning Improperly  Left Front Bumper,Center Front End,Left Front …
2           Unspecified                          Right Rear Quarter Panel
3           Unspecified                          Left Front Quarter Panel
4           Unspecified  Right Rear Bumper,Center Back End,Left Rear Bu…

[5 rows x 22 columns]
```

2.0.6 COUNT

Count of collisions for each jurisdiction.

[11]:
```
#pd.set_option('display.max_rows', None)
license_jurisdiction_counts = data.groupby('DRIVER_LICENSE_JURISDICTION').
 ↪size().reset_index(name='COLLISION_COUNT')
license_jurisdiction_counts.sort_values(by='COLLISION_COUNT', ascending =␣
 ↪False).head().reset_index(drop=True)
```

[11]:
```
  DRIVER_LICENSE_JURISDICTION  COLLISION_COUNT
0                          NY           130622
1                          NJ             9284
2                          PA             2524
```

```
3                          CT              1658
4                          FL              1649
```

Count of Travel directions in the Dataset

[12]:
```
travel_direction_counts = data.groupby('TRAVEL_DIRECTION').size().
 ↪reset_index(name='COLLISION_COUNT')
travel_direction_counts
```

[12]:

	TRAVEL_DIRECTION	COLLISION_COUNT
0	-	156
1	E	62
2	East	35708
3	N	71
4	North	35357
5	Northeast	2619
6	Northwest	2378
7	S	82
8	South	34835
9	Southeast	2456
10	Southwest	2355
11	U	4
12	Unknown	2053
13	W	81
14	West	36038

Transform different values into categorical for varibale '### Convert 'VEHICLE_YEAR' data type to int format'.

[13]:
```
data['TRAVEL_DIRECTION'] = data['TRAVEL_DIRECTION'].replace({
    'E': 'East',
    'N': 'North',
    'S':'South',
    '-':'Unknown',
    'U':'Unknown',
    'W':'West'
})
travel_direction_counts = data.groupby('TRAVEL_DIRECTION').size().
 ↪reset_index(name='COLLISION_COUNT')
travel_direction_counts.sort_values(by='COLLISION_COUNT', ascending = False)
```

[13]:

	TRAVEL_DIRECTION	COLLISION_COUNT
8	West	36119
0	East	35770
1	North	35428
4	South	34917
2	Northeast	2619
5	Southeast	2456

```
3        Northwest             2378
6        Southwest             2355
7          Unknown             2213
```

Count of Points of Impact in the Dataset : There are 18 point of impact in vehicle that's categorized.

[14]:
```
point_of_impact_counts = data.groupby('POINT_OF_IMPACT').size().
 ↪reset_index(name='COLLISION_COUNT')
point_of_impact_counts = point_of_impact_counts.
 ↪sort_values(by='COLLISION_COUNT', ascending =False).reset_index(drop=True)
point_of_impact_counts
# Center Front End has highest damages in vehicles through accidents in NY␣
 ↪state.
```

[14]:

	POINT_OF_IMPACT	COLLISION_COUNT
0	Center Front End	24970
1	Center Back End	19864
2	Left Front Bumper	19374
3	Right Front Bumper	18116
4	Right Front Quarter Panel	12268
5	Left Front Quarter Panel	10553
6	Left Rear Quarter Panel	7964
7	Left Side Doors	7879
8	Right Side Doors	7854
9	Left Rear Bumper	7277
10	Right Rear Quarter Panel	6951
11	Right Rear Bumper	5980
12	No Damage	2943
13	Other	1424
14	Trailer	355
15	Roof	211
16	Undercarriage	116
17	Overturned	97
18	Demolished	65

Licensed people have made more accidents

[15]:
```
license_status_counts = data.groupby('DRIVER_LICENSE_STATUS').size().
 ↪reset_index(name='COLLISION_COUNT')
license_status_counts = license_status_counts.sort_values(by='COLLISION_COUNT',␣
 ↪ascending =False).reset_index(drop=True)
license_status_counts
# Licensed people have made more accidents
```

[15]:

	DRIVER_LICENSE_STATUS	COLLISION_COUNT
0	Licensed	150661

```
1            Unlicensed            2594
2                Permit            1016
```

summarizing the counts of collisions based on the status of the car just before crashing

[16]:
```
pre_crash_counts = data.groupby('PRE_CRASH').size().
↪reset_index(name='COLLISION_COUNT')
pre_crash_counts = pre_crash_counts.sort_values(by='COLLISION_COUNT', ascending␣
↪=False).reset_index(drop=True)
pre_crash_counts
```

[16]:

	PRE_CRASH	COLLISION_COUNT
0	Going Straight Ahead	83444
1	Making Left Turn	10867
2	Making Right Turn	9104
3	Stopped in Traffic	9083
4	Parked	8665
5	Slowing or Stopping	7044
6	Backing	6027
7	Changing Lanes	5528
8	Merging	3181
9	Starting from Parking	2975
10	Entering Parked Position	2396
11	Passing	1621
12	Making U Turn	1398
13	Other*	1165
14	Starting in Traffic	786
15	Avoiding Object in Roadway	186
16	Making Right Turn on Red	37
17	Making Left Turn on Red	36
18	Police Pursuit	32

[17]:
```
data.head()
```

[17]:

	UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE
0	17307404	3493857	2012-07-31	13:30	591401	NY	Box Truck	GMC-TRUCK/BUS
1	17295384	3518768	2012-09-12	12:20	780983	NY	Sedan	HOND -CAR/SUV
2	17218669	3576809	2012-12-09	12:03	1178874	NJ	Taxi	CADI -CAR/SUV
3	17143961	3421148	2013-04-05	09:40	453654	NY	Station Wagon/Sport Utility Vehicle	FORD -CAR/SUV
4	17272443	3467155	2013-06-23	08:30	437750	NY	4 dr sedan	TOYT -CAR/SUV

```
  VEHICLE_MODEL  VEHICLE_YEAR  … DRIVER_SEX  DRIVER_LICENSE_STATUS  \
0           NaN          1997  …          M                Licensed
1           NaN          2000  …          M                Licensed
2           NaN          2015  …          M                Licensed
3      FORD ECP          2015  …          M                Licensed
4           NaN          2012  …          F                Licensed

  DRIVER_LICENSE_JURISDICTION                PRE_CRASH  \
0                          NY            Changing Lanes
1                          NY          Making Left Turn
2                          NJ  Entering Parked Position
3                          NY                  Backing
4                          NY      Going Straight Ahead

           POINT_OF_IMPACT PUBLIC_PROPERTY_DAMAGE  \
0          Right Side Doors                      N
1         Left Front Bumper                      N
2  Right Rear Quarter Panel                      N
3  Left Front Quarter Panel                      N
4         Right Rear Bumper                      N

  PUBLIC_PROPERTY_DAMAGE_TYPE           CONTRIBUTING_FACTOR_1  \
0                         NaN  Driver Inattention/Distraction
1                         NaN             Driver Inexperience
2                         NaN             Passing Too Closely
3                         NaN                     Unspecified
4                         NaN                     Unspecified

   CONTRIBUTING_FACTOR_2                                  VEHICLE_DAMAGES
0           Unspecified                                     Right Side Doors
1     Turning Improperly  Left Front Bumper,Center Front End,Left Front …
2           Unspecified                          Right Rear Quarter Panel
3           Unspecified                          Left Front Quarter Panel
4           Unspecified  Right Rear Bumper,Center Back End,Left Rear Bu…

[5 rows x 22 columns]
```

PUBLIC_PROPERTY_DAMAGE are divided in 3 categories with Yes / No / Unspecified

[18]:
```
public_property_damage_counts = data.groupby('PUBLIC_PROPERTY_DAMAGE').size().
↪reset_index(name='COLLISION_COUNT')
public_property_damage_counts = public_property_damage_counts.
↪sort_values(by='COLLISION_COUNT', ascending =False).reset_index(drop=True)
public_property_damage_counts
```

[18]:
```
  PUBLIC_PROPERTY_DAMAGE  COLLISION_COUNT
0                      N          146177
1            Unspecified            7368
2                      Y             726
```

'PUBLIC_PROPERTY_DAMAGE_TYPE' are just filled values Nan, Hence removed the column that does'nt provide exta information.

[19]:
```
public_property_damage_type_counts = data.
 ↪groupby('PUBLIC_PROPERTY_DAMAGE_TYPE').size().
 ↪reset_index(name='COLLISION_COUNT')
public_property_damage_type_counts = public_property_damage_type_counts.
 ↪sort_values(by='COLLISION_COUNT', ascending =False).reset_index(drop=True)
public_property_damage_type_counts
```

[19]:
```
Empty DataFrame
Columns: [PUBLIC_PROPERTY_DAMAGE_TYPE, COLLISION_COUNT]
Index: []
```

[20]:
```
data.drop(columns=['PUBLIC_PROPERTY_DAMAGE_TYPE'], inplace=True)
```

[21]:
```
contributing_factor_1_counts = data.groupby('CONTRIBUTING_FACTOR_1').size().
 ↪reset_index(name='COLLISION_COUNT')
contributing_factor_1_counts = contributing_factor_1_counts.
 ↪sort_values(by='COLLISION_COUNT', ascending =False).reset_index(drop=True)
contributing_factor_1_counts
```

[21]:
```
             CONTRIBUTING_FACTOR_1  COLLISION_COUNT
0                      Unspecified            76440
1   Driver Inattention/Distraction            24380
2            Following Too Closely            10058
3    Failure to Yield Right-of-Way             6658
4   Passing or Lane Usage Improper             4303
..                             …              …
51         Cell Phone (hands-free)                7
52    Shoulders Defective/Improper                5
53            Windshield Inadequate                5
54                          Texting                4
55                Vehicle Vandalism                4

[56 rows x 2 columns]
```

[22]:
```
contributing_factor_2_counts = data.groupby('CONTRIBUTING_FACTOR_2').size().
 ↪reset_index(name='COLLISION_COUNT')
contributing_factor_2_counts = contributing_factor_2_counts.
 ↪sort_values(by='COLLISION_COUNT', ascending =False).reset_index(drop=True)
contributing_factor_2_counts
```

[22]:
```
              CONTRIBUTING_FACTOR_2  COLLISION_COUNT
0                       Unspecified           123662
1   Driver Inattention/Distraction              7103
2            Following Too Closely              3462
3    Failure to Yield Right-of-Way              2405
4              Unsafe Lane Changing             1813
..                              …                …
50             Headlights Defective                3
51           Other Lighting Defects                2
52    Shoulders Defective/Improper                 2
53      Listening/Using Headphones                 2
54            Windshield Inadequate                1

[55 rows x 2 columns]
```

2.0.7 IS NULL

[23]:
```
data['VEHICLE_TYPE'].isnull().sum()
```

[23]:
```
4
```

There are collision ID that are duplicated : mean multiple vehicle involved in single collision

[24]:
```
data["COLLISION_ID"].duplicated().sum()
```

[24]:
```
37092
```

Aggregate vehicle IDs by Collision ID : Max no. of vehicles involved in single collision is 6.

[25]:
```
collision_vehicle_count = data.groupby("COLLISION_ID").agg({
    "VEHICLE_ID": "count",
}).reset_index()
unique_car_counts = collision_vehicle_count["VEHICLE_ID"].nunique()
sort_collision_vehicle_count = collision_vehicle_count.
 ↪sort_values(by="VEHICLE_ID", ascending=False).reset_index(drop=True)
sort_collision_vehicle_count
```

[25]:
```
        COLLISION_ID  VEHICLE_ID
0            3501692           6
1            3454136           6
2            3434859           6
3            3478656           5
4            3445871           5
…              …         …
117174       3480815           1
117175       3480812           1
```

```
117176       3480809           1
117177       3480804           1
117178       4466944           1

[117179 rows x 2 columns]
```

[26]:
```
collision_vehicle_count = sort_collision_vehicle_count.
↪groupby("VEHICLE_ID")["COLLISION_ID"].count().reset_index()
collision_vehicle_count = collision_vehicle_count.sort_values(by="VEHICLE_ID",␣
↪ascending=False).reset_index(drop=True)
collision_vehicle_count
# There are 3 times collision happened that resulted in 6 vehicles crash at␣
↪same time.
```

[26]:

	VEHICLE_ID	COLLISION_ID
0	6	3
1	5	10
2	4	98
3	3	1013
4	2	34717
5	1	81338

2.0.8 COUNTS

[27]:
```
vehicle_type_counts = data['VEHICLE_TYPE'].value_counts().
↪reset_index(name='COLLISION_COUNT')
vehicle_type_counts = vehicle_type_counts.rename(columns={'index':␣
↪'VEHICLE_TYPE'})
vehicle_type_counts
```

[27]:

	VEHICLE_TYPE	COLLISION_COUNT
0	Station Wagon/Sport Utility Vehicle	48000
1	Sedan	44759
2	4 dr sedan	31315
3	Taxi	6921
4	Pick-up Truck	4444
..	…	…
402	tow t	1
403	wg	1
404	DODGE	1
405	renta	1
406	SCOOTER	1

[407 rows x 2 columns]

[28]:
```
pd.set_option('display.max_columns', None)
data.head()
```

[28]:

	UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID
0	17307404	3493857	2012-07-31	13:30	591401
1	17295384	3518768	2012-09-12	12:20	780983
2	17218669	3576809	2012-12-09	12:03	1178874
3	17143961	3421148	2013-04-05	09:40	453654
4	17272443	3467155	2013-06-23	08:30	437750

	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE
0	NY	Box Truck	GMC-TRUCK/BUS
1	NY	Sedan	HOND -CAR/SUV
2	NJ	Taxi	CADI -CAR/SUV
3	NY	Station Wagon/Sport Utility Vehicle	FORD -CAR/SUV
4	NY	4 dr sedan	TOYT -CAR/SUV

	VEHICLE_MODEL	VEHICLE_YEAR	TRAVEL_DIRECTION	VEHICLE_OCCUPANTS	DRIVER_SEX
0	NaN	1997	West	0	M
1	NaN	2000	South	1	M
2	NaN	2015	West	0	M
3	FORD ECP	2015	North	1	M
4	NaN	2012	South	2	F

	DRIVER_LICENSE_STATUS	DRIVER_LICENSE_JURISDICTION	PRE_CRASH
0	Licensed	NY	Changing Lanes
1	Licensed	NY	Making Left Turn
2	Licensed	NJ	Entering Parked Position
3	Licensed	NY	Backing
4	Licensed	NY	Going Straight Ahead

	POINT_OF_IMPACT	PUBLIC_PROPERTY_DAMAGE
0	Right Side Doors	N
1	Left Front Bumper	N
2	Right Rear Quarter Panel	N
3	Left Front Quarter Panel	N
4	Right Rear Bumper	N

	CONTRIBUTING_FACTOR_1	CONTRIBUTING_FACTOR_2
0	Driver Inattention/Distraction	Unspecified
1	Driver Inexperience	Turning Improperly
2	Passing Too Closely	Unspecified
3	Unspecified	Unspecified
4	Unspecified	Unspecified

	VEHICLE_DAMAGES
0	Right Side Doors
1	Left Front Bumper,Center Front End,Left Front …
2	Right Rear Quarter Panel
3	Left Front Quarter Panel

```
4  Right Rear Bumper,Center Back End,Left Rear Bu…
```

2.0.9 NULL VALUES CONTAINING VARIABLES

[95]:
```
data['STATE_REGISTRATION'].isnull().sum()
```

[95]:
```
801
```

[99]:
```
data['VEHICLE_TYPE'].isnull().sum()
```

[99]:
```
4
```

[100]:
```
data['VEHICLE_MAKE'].isnull().sum()
```

[100]:
```
3011
```

[101]:
```
data['VEHICLE_MODEL'].isnull().sum()
```

[101]:
```
132180
```

[102]:
```
data['TRAVEL_DIRECTION'].isnull().sum()
```

[102]:
```
16
```

[109]:
```
data['DRIVER_LICENSE_JURISDICTION'].isnull().sum()
```

[109]:
```
1608
```

[110]:
```
data['PRE_CRASH'].isnull().sum()
```

[110]:
```
696
```

[111]:
```
data['POINT_OF_IMPACT'].isnull().sum()
```

[111]:
```
10
```

[113]:
```
data['CONTRIBUTING_FACTOR_1'].isnull().sum()
```

[113]:
```
15
```

[114]:
```
data['CONTRIBUTING_FACTOR_2'].isnull().sum()
```

[114]:
```
2113
```

2.0.10 Step 4: Meeting Schedule

We are all busy. You and your team should be working towards this project on a weekly basis. Share your proposed schedule below.

For last 2 weeks we are usually meeting after the class on Friday Morning EST to heads up on the project tasks to be accomplished and divide tasks accordingly.

On Monday Morning at flexible timings we sit togethor to discuss everyone's input and accomplish the codes and it's comments.

1. Tasks divided
   a.Work on Research questions.
   b.check which variables are effective for answering effective questions and analysis.
   c.Learn and paerform different techniques to clean and transform the data.

[ ]: