MGT 6203 Group Project Proposal Template

Please edit the following template to record your responses and provide details on your project plan.

TEAM INFORMATION (1 point)

Team #: 45

Team Members:

1. Jeff Hedberg; jhedberg3@gatech.edu (Team Lead)
   Data Scientist for John Deere (20+ years) with an undergraduate degree in Mechanical Engineering and an MBA. Previous analytic work patented (9,765,690) for predicting engine failure on agriculture and construction equipment using telematic data. Current work involves developing and productionizing ML models for supply chain operations.
2. Lisa Chille; lchille@gatech.edu
   …
3. Brittany Lange; blange9@gatech.edu
   …
4. Delband Taha; dtaha@gatech.edu
   …
5. Nick Cunningham; ncunningham8@gatech.edu
   …

OBJECTIVE/PROBLEM (5 points)

<u>Project Title:</u>

Predicting NYC Auto Collision Volume for a local repair shop

<u>Background Information on chosen project topic:</u>

NYC has made public several datasets on vehicle collisions. This project aims at leveraging those datasets to help a local auto repair center estimate the volume of incoming vehicles they can expect in the future year (given their known market share). Having this data, the company can better plan for staffing levels in advance of needing to repair the vehicles.

We will also look at the best demographic for the business to target in order to increase market share. Through this process we can suggest some incremental share growth and then compute the incremental volume that they can expect as a result.

If we have extra time, we will also will briefly explore the impact of COVID-19 on this industry, in recent years.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

Use historic auto collision data from NYC to estimate the future collision volume. Then, take these collision estimates and combine them with a known market share of a local auto repair business to forecast their expected auto repair volume. Additionally, recommend a marketing campaign for market share growth and then forecast the expected incremental gains if the campaign is successful (by a known growth percent per demographic). Additionally, if time permits, evaluate the effect of Covid-19 on this industry sector.

State your Primary Research Question (RQ):

1. What is the historic volume of NYC Auto Collisions?
2. Can this historic volume of NYC Auto Collisions be used to accurately forecast Auto Collision Volumes in a future year?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Can the Auto Collision Volume estimates be extended to local repair shop volumes?
2. What demographic should be targeted for a business wanting to grow their Auto Collision repair market share?
3. How much would a successful marketing campaign increase Auto Collision Volume?
4. Did Covid-19 change the overall Auto Collision Volume in NYC?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

An owner of a local Auto Collision Repair business would be very interested in forecasting the volume of cars that would require service. This would allow for them to better plan staffing levels for the future months and year. In addition, if this could help them determine if the facilities, they are currently using are large enough to support the volume of business without losing market share. Additionally, if they have ambitions for market share growth, they would like to know which demographic to target and an understanding of how much a gin in market share translates into addition auto collision volume for their shop.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):
The main data sources for this project are sourced from Data.GOV

- NYC Motor Vehicle Collisions – Crashes: Link — Rows: 1,896,229 — Columns:29
- NYC Motor Vehicle Collisions – Vehicles: Link — Rows: 4,692,054 — Columns:21
- NYC Motor Vehicle Collisions – Person: Link — Rows: 3,704,406 — Columns:25

Data Description (describe each of your data sources, include screenshots of a few rows of data):

The Crashes dataset contains data on the actual crash events in NYC from 2013-01-01 to 2021-12-31.

Below are several screenshots of the data since it’s so wide:

CRASH.DATE	CRASH.TIME	BOROUGH	ZIP.CODE	LATITUDE	LONGITUDE	LOCATION	ON.STREET.NAME	CROSS.STREET.NAME	OFF.STREET.NAME	NUMBER.OF.PERSONS.INJURED	NUMBER.OF.PERSONS.KILLED
4/14/2021	5:32		NA	NA	NA		BRONX WHITESTONE BRIDGE			0	0
4/13/2021	21:35	BROOKLYN	11217	40.68358	-73.97617	(40.68358, -73.97617)			620 ATLANTIC AVENUE	1	0
4/15/2021	16:15		NA	NA	NA		HUTCHINSON RIVER PARKWAY			0	0
4/13/2021	16:00	BROOKLYN	11222	NA	NA		VANDERVORT AVENUE	ANTHONY STREET		0	0
4/12/2021	8:25		NA	0	0	(0.0, 0.0)	EDSON AVENUE			0	0
4/13/2021	17:11		NA	NA	NA		VERRAZANO BRIDGE UPPER			0	0

NUMBER.OF.PEDESTRIANS.INJURED	NUMBER.OF.PEDESTRIANS.KILLED	NUMBER.OF.CYCLIST.INJURED	NUMBER.OF.CYCLIST.KILLED	NUMBER.OF.MOTORIST.INJURED	NUMBER.OF.MOTORIST.KILLED	CONTRIBUTING.FACTOR.VEHICLE.1	CONTRIBUTING.FACTOR.VEHICLE.2
0	0	0	0	0	0	Following Too Closely	Unspecified
1	0	0	0	0	0	Unspecified	
0	0	0	0	0	0	Pavement Slippery	
0	0	0	0	0	0	Following Too Closely	Unspecified
0	0	0	0	0	0	Unspecified	Unspecified
0	0	0	0	0	0	Following Too Closely	Unspecified

CONTRIBUTING.FACTOR.VEHICLE.3	CONTRIBUTING.FACTOR.VEHICLE.4	CONTRIBUTING.FACTOR.VEHICLE.5	COLLISION_ID	VEHICLE.TYPE.CODE.1	VEHICLE.TYPE.CODE.2	VEHICLE.TYPE.CODE.3	VEHICLE.TYPE.CODE.4	VEHICLE.TYPE.CODE.5
			4407480	Sedan	Sedan			
			4407147	Sedan				
			4407665	Station Wagon/Sport Utility Vehicle				
			4407811	Sedan				
			4406885	Station Wagon/Sport Utility Vehicle	Sedan			
			4407883	Sedan	Box Truck			

The vehicle dataset contains data on the actual crash vehicles in NYC from 2013-01-01 to 2021-12-31.

Below are several screenshots of the data since it’s so wide:

UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	VEHICLE_ID	STATE_REGISTRATION	VEHICLE_TYPE	VEHICLE_MAKE	VEHICLE_MODEL	VEHICLE_YEAR	TRAVEL_DIRECTION	VEHICLE_OCCUPANTS	DRIVER_SEX	DRIVER_LICENSE_STATUS
10385780	100201	9/7/2012	9:03	1	NY	PASSENGER VEHICLE			NA		NA		
19140702	4213082	9/23/2019	8:15	0553ab4d-9500-4cba-8d98-f4d7f89d5856	NY	Station Wagon/Sport Utility Vehicle	TOYT -CAR/SUV		2002	North	1	M	Licensed
14887647	3307608	10/2/2015	17:18	2	NY	TAXI			NA		NA		
14889754	3308693	10/4/2015	20:34	1	NY	PASSENGER VEHICLE			NA		NA		
14400270	297666	4/25/2013	21:15	1	NY	PASSENGER VEHICLE			NA		NA		
17044639	3434155	5/2/2016	17:35	219456	NY	4 dr sedan	MERZ -CAR/SUV		2015	East	2	M	Licensed

DRIVER_LICENSE_JURISDICTION	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
									Unspecified	
NY	Going Straight Ahead	Left Front Bumper	Left Front Quarter Panel				N		Driver Inattention/Distraction	Unspecified
	Going Straight Ahead								Driver Inattention/Distraction	
	Parked								Unspecified	
									Other Vehicular	
FL	Merging	Right Front Bumper	Right Front Bumper	Right Front Quarter Panel			N		Driver Inattention/Distraction	Unsafe Lane Changing

The person dataset contains data on the actual crash persons in NYC from 2013-01-01 to 2021-12-31.

Below are several screenshots of the data since it’s so wide:

UNIQUE_ID	COLLISION_ID	CRASH_DATE	CRASH_TIME	PERSON_ID	PERSON_TYPE	PERSON_INJURY	VEHICLE_ID	PERSON_AGE	EJECTION	EMOTIONAL_STATUS	BODILY_INJURY
10249006	4229554	10/26/2019	9:43	31aa2bc0-f545-444f-8cdb-f1cb5cf00b89	Occupant	Unspecified	19141108	NA			
10255054	4230587	10/25/2019	15:15	4629e500-a73e-48dc-b8fb-53124d124b80	Occupant	Unspecified	19144075	33	Not Ejected	Does Not Apply	Does Not Apply
10253177	4230550	10/26/2019	17:55	ae48c136-1383-45db-83f4-2a5eecfb7cff	Occupant	Unspecified	19143133	55			
6650180	3565527	11/21/2016	13:05	2782525	Occupant	Unspecified	NA	NA			
10255516	4231168	10/25/2019	11:16	e038e18f-40fb-4471-99cf-345eae36e064	Occupant	Unspecified	19144329	7	Not Ejected	Does Not Apply	Does Not Apply
10253606	4230743	10/24/2019	19:15	84bcb3a7-d201-4c61-9e30-fe29268c1074	Occupant	Injured	19143343	27	Not Ejected	Conscious	Back

POSITION_IN_VEHICLE	SAFETY_EQUIPMENT	PED_LOCATION	PED_ACTION	COMPLAINT	PED_ROLE	CONTRIBUTING_FACTOR_1	CONTRIBUTING_FACTOR_2	PERSON_SEX
					Registrant			U
Front passenger, if two or more persons, including the driver, are in the front seat	Lap Belt & Harness			Does Not Apply	Passenger			F
					Registrant			M
					Notified Person			
Right rear passenger or motorcycle sidecar passenger	Lap Belt			Does Not Apply	Passenger			F
Driver	Lap Belt & Harness			Complaint of Pain or Nausea	Driver			M

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Key Dependent Variable:

1. Volume of Auto Collisions – This will need to be computed historically from the individual records

Key Independent Variables:

1. CRASH_DATE – This will be used to create new time related features (Month, Year, Year_Month, etc.)
2. PERSON_AGE – Useful for modeling demographic factors
3. PERSON_SEX – Useful for modeling demographic factors
4. PERSON_TYPE – Needed to filter for just 'Drivers' (to get correct collision counts)
5. TIMESTEP – Need to create this feature as a index of Year_Month. This will be used to tie all months and years together in chronological order for modeling.
6. AGE_GROUP – Need to create this feature for demographic modeling. This will be a binned version of the PERSON_AGE field (likely 5 year increments from 15-100).

Above is a list of the variables that we expect to be most relevant when solving the problem we've outlined. Many of these will need to be created from other data elements in our source data. Additionally, there are categorical features here that will need to be represented as different factor levels when modeling.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

Initially, we will need to join all independent datasets into a common combined dataset. Then, we will need to engineer the appropriate features as outlined above for Auto Collision Volume Modeling. Next, we will need to aggregate the data to a Year_Month granularity with partitioning for all demographic features. This will ensure that we can model each demographic impact on overall Auto Collision Volume at a Year_Month level. Prior to performing the modeling we'll need to create appropriate train/test splits. Since we are dealing with a temporal model, we will most likely select all except the final year for training, and then the final year for testing. This will allow us to build many different model types (multiple linear, random forest, CART tree, boosted tree, etc.) and then compare their relative performances. After selecting a final model (or ensemble) we will then compute the Auto Collision Volume for the local repair shop (using their market share). Additional project work will then focus on suggesting a target demographic for market share growth, followed by computing the impact in additional Auto Collision Volume for a successful campaign.

If we have time, we can also look at the impact of Covid-19 on the Auto Collision Volume and understand if any additional strains were placed on this industry sector.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

I expect that we will be able build, validate, and utilize a ML model for predicting Auto Collision Volumes in the future year for a local Auto Repair Shop in NYC. I also believe that we will be able to estimate their monthly Auto Collision Volume with good accuracy. In addition, I believe that we will be able to identify a key customer demographic for them to target with a market share growth campaign, and that we'll be able to quantify that market share growth in terms of additional Auto Collision Volume using our model.

I believe that we'll find that Covid-19 impacted this industry sector volumes since the lockdowns mandated residents stay at home, and many employers switched to a remote employee base.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

As previously mentioned, our project solution can help a local NYC Auto Collision Repair Shop determine appropriate staffing levels for the future year for their business. It can also help them evaluate if they need to increase or decrease their current operational footprint (facility sizes, market share, etc.). We can also help them target specific demographics for best potential of market share growth, as well as quantify that growth into estimated Auto Collision Volume increases. This should be a very useful analytic product for any business in that NYC Auto Collision Repair industry sector.

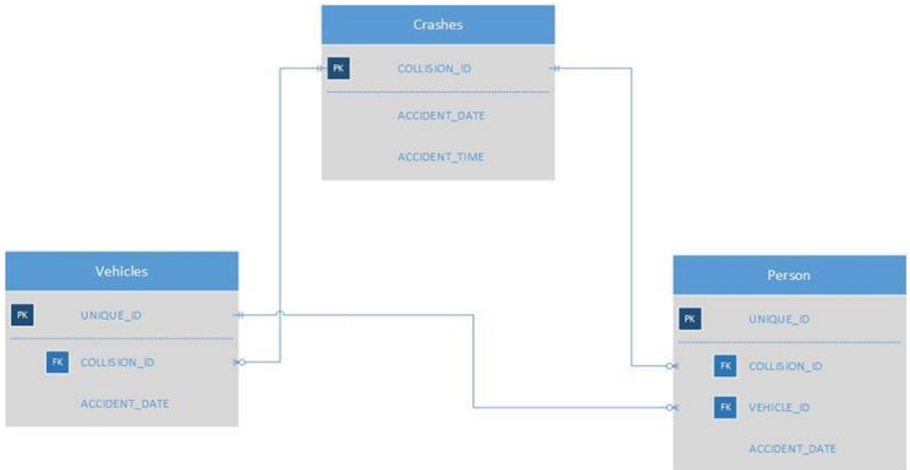
PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

- 2022-06-24 – Initial EDA Complete
- 2022-06-24 – Start on Data Transformations
- 2022-06-30 – Progress Video Complete
- 2022-06-30 – Start on model Building
- 2022-07-04 – Progress Report Complete
- 2022-07-12 – Project Findings Complete (All modeling and Demographic Analysis)
- 2022-07-12 – Project Additional Findings Complete (Impact of Covid-19)
- 2022-07-15 – Final Video Complete
- 2022-07-20 – Final Report Complete

Appendix (any preliminary figures or charts that you would like to include):

Below is a n Entity Relationship Diagram from the NYC Dataset documentation:



Below are descriptive statistics for the 3 datasets that we plan to use in this project:

Crashes Dataset

```r
#### Load Data
crashes_df <- read.csv('./Motor_Vehicle_Collisions_-_Crashes.csv', stringsAsFactors = FALSE) %>%
  mutate(CRASH.DATE = as.Date(CRASH.DATE, "%m/%d/%Y")) #1,896,229 x 29

# crashes_df
# min(crashes_df$CRASH.DATE) #"2012-07-01"
# max(crashes_df$CRASH.DATE) #"2022-05-29"

kable(t(summary(crashes_df))) %>% kable_classic(full_width = TRUE, html_font = "Cambria", font_size = 14)
```

CRASH.DATE	Min. :2012-07-01	1st Qu.:2014-10-28	Median :2016-12-15	Mean :2017-01-01	3rd Qu.:2019-01-04	Max. :2022-05-29	NA
CRASH.TIME	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
BOROUGH	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
ZIP.CODE	Min. :10000	1st Qu.:10306	Median :11207	Mean :10837	3rd Qu.:11237	Max. :11697	NA's :587695
LATITUDE	Min. : 0.00	1st Qu.:40.67	Median :40.72	Mean :40.64	3rd Qu.:40.77	Max. :43.34	NA's :220042
LONGITUDE	Min. :-201.36	1st Qu.: -73.98	Median : -73.93	Mean : -73.77	3rd Qu.: -73.87	Max. : 0.00	NA's :220042
LOCATION	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
ON.STREET.NAME	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
CROSS.STREET.NAME	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
OFF.STREET.NAME	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
NUMBER.OF.PERSONS.INJURED	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.2873	3rd Qu.: 0.0000	Max. :43.0000	NA's :18
NUMBER.OF.PERSONS.KILLED	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.001358	3rd Qu.:0.000000	Max. :8.000000	NA's :31
NUMBER.OF.PEDESTRIANS.INJURED	Min. : 0.00000	1st Qu.: 0.00000	Median : 0.00000	Mean : 0.05304	3rd Qu.: 0.00000	Max. :27.00000	NA
NUMBER.OF.PEDESTRIANS.KILLED	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.000697	3rd Qu.:0.000000	Max. :6.000000	NA
NUMBER.OF.CYCLIST.INJURED	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.02435	3rd Qu.:0.00000	Max. :4.00000	NA
NUMBER.OF.CYCLIST.KILLED	Min. :0.0000000	1st Qu.:0.0000000	Median :0.0000000	Mean :0.0001007	3rd Qu.:0.0000000	Max. :2.0000000	NA
NUMBER.OF.MOTORIST.INJURED	Min. : 0.0000	1st Qu.: 0.0000	Median : 0.0000	Mean : 0.2083	3rd Qu.: 0.0000	Max. :43.0000	NA
NUMBER.OF.MOTORIST.KILLED	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.00055	3rd Qu.:0.00000	Max. :5.00000	NA
CONTRIBUTING.FACTOR.VEHICLE.1	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING.FACTOR.VEHICLE.2	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING.FACTOR.VEHICLE.3	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING.FACTOR.VEHICLE.4	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING.FACTOR.VEHICLE.5	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
COLLISION_ID	Min. : 22	1st Qu.:3046695	Median :3584305	Mean :3021392	3rd Qu.:4058626	Max. :4533068	NA
VEHICLE.TYPE.CODE.1	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE.TYPE.CODE.2	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE.TYPE.CODE.3	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE.TYPE.CODE.4	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE.TYPE.CODE.5	Length:1896229	Class :character	Mode :character	NA	NA	NA	NA

Person Dataset

```
#### Load Data
person_df <- read.csv('./Motor_Vehicle_Collisions_-_Person.csv', stringsAsFactors = FALSE) %>%
  mutate(CRASH_DATE = as.Date(CRASH_DATE, "%m/%d/%Y")) #4,692,054 x 21

# person_df
# min(person_df$CRASH_DATE) #"2012-07-01"
# max(person_df$CRASH_DATE) #"2022-05-29"

kable(t(summary(person_df))) %>% kable_classic(full_width = TRUE, html_font = "Cambria", font_size = 14)
```

UNIQUE_ID	Min. : 10922	1st Qu.: 6812186	Median : 8996148	Mean : 8531863	3rd Qu.:10216281	Max. :12239058	NA
COLLISION_ID	Min. : 37	1st Qu.:3638855	Median :3921823	Mean :3853306	3rd Qu.:4210666	Max. :4533068	NA
CRASH_DATE	Min. :2012-07-01	1st Qu.:2017-03-19	Median :2018-06-08	Mean :2018-07-08	3rd Qu.:2019-09-20	Max. :2022-05-29	NA
CRASH_TIME	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PERSON_ID	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PERSON_TYPE	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PERSON_INJURY	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_ID	Min. : 123423	1st Qu.:17466247	Median :18528882	Mean :18253620	3rd Qu.:19125401	Max. :20229580	NA's :185684
PERSON_AGE	Min. :-999.0	1st Qu.: 23.0	Median : 35.0	Mean : 36.8	3rd Qu.: 50.0	Max. :9999.0	NA's :453265
EJECTION	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
EMOTIONAL_STATUS	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
BODILY_INJURY	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
POSITION_IN_VEHICLE	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
SAFETY_EQUIPMENT	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PED_LOCATION	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PED_ACTION	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
COMPLAINT	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PED_ROLE	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING_FACTOR_1	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING_FACTOR_2	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA
PERSON_SEX	Length:4692054	Class :character	Mode :character	NA	NA	NA	NA

Vehicle Dataset

```
#### Load Data
vehicles_df <- read.csv('./Motor_Vehicle_Collisions_-_Vehicles.csv', stringsAsFactors = FALSE) %>%
  mutate(CRASH_DATE = as.Date(CRASH_DATE, "%m/%d/%Y")) #3,704,406 x 25

# vehicles_df
# min(vehicles_df$CRASH_DATE) #"2012-07-01"
# max(vehicles_df$CRASH_DATE) #"2021-12-04"

kable(t(summary(vehicles_df))) %>% kable_classic(full_width = TRUE, html_font = "Cambria", font_size = 14)
```

UNIQUE_ID	Min. : 111711	1st Qu.:14215160	Median :17306058	Mean :16060871	3rd Qu.:18739205	Max. :20121717	NA
COLLISION_ID	Min. : 22	1st Qu.:3017853	Median :3567068	Mean :2996659	3rd Qu.:4028145	Max. :4484197	NA
CRASH_DATE	Min. :2012-07-01	1st Qu.:2014-10-15	Median :2016-11-18	Mean :2016-11-21	3rd Qu.:2018-11-15	Max. :2021-12-04	NA
CRASH_TIME	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_ID	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
STATE_REGISTRATION	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_TYPE	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_MAKE	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_MODEL	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_YEAR	Min. : 1000	1st Qu.: 2008	Median : 2013	Mean : 2015	3rd Qu.: 2016	Max. :20063	NA's :1796971
TRAVEL_DIRECTION	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_OCCUPANTS	Min. :0.00e+00	1st Qu.:1.00e+00	Median :1.00e+00	Mean :1.01e+03	3rd Qu.:1.00e+00	Max. :1.00e+09	NA's :1718406
DRIVER_SEX	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
DRIVER_LICENSE_STATUS	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
DRIVER_LICENSE_JURISDICTION	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
PRE_CRASH	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
POINT_OF_IMPACT	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_DAMAGE	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_DAMAGE_1	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_DAMAGE_2	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
VEHICLE_DAMAGE_3	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
PUBLIC_PROPERTY_DAMAGE	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
PUBLIC_PROPERTY_DAMAGE_TYPE	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING_FACTOR_1	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA
CONTRIBUTING_FACTOR_2	Length:3704406	Class :character	Mode :character	NA	NA	NA	NA