Project Proposal

Team Information

Team Number

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
   I am a full-stack Software Engineer at Dropbox. I have experience working on planet-scale (110+ million daily active users) cloud services and infrastructure from Microsoft (Windows, Office, Cloud PC). I graduated with a Computer Science degree from Harvard University in 2018. My analytics projects background comprises of ISyE 6501 final project and assignments.

3. **Brittany Lange**; blange9@gatech.edu
   I have a BS in Mathematics from Cal Poly and I have spent the last several years as a stay at home parent to my kids. Now that my kids will all be in school this fall, I decided to change careers and enter the world of data analysis, which I have found a passion for. I do not have much experience yet, but I am learning more every day!

4. **Delband Taha**; dtaha@gatech.edu
   I am a data analyst working for the local school district with and graduated from Virginia Tech with degrees in Accounting and Finance. Previous analytics work at Booz Allen Hamilton led to a slew of changes to a few policies, the onboarding and training, and the disciplinary process.

5. **Nick Cunningham**; ncunningham8@gatech.edu

   I am a Senior Director at Gap Inc. and lead a team of business intelligence developers, digital analysts and project managers. I am also the Chief of Staff to the SVP of Data Science and Analytics. I have a BA in Political Science from the University of Colorado and completed a full-time Data Science immersive. Most of my work is leading strategic projects and team management in the retail analytics space, but the GT program has allowed me to continue to cultivate hands on experience.

Objective/Problem

Title

Predicting NYC Auto Collision Volume for a local repair shop.

Background Information

While motor vehicle collisions have emotional, financial and health costs associated with them, they also present an opportunity to add/restore value in the form of auto repairs. This project aims to help a fictitious[1] local auto repair center in New York state, estimate the volume of incoming vehicles they can expect each year, assuming their market share[2] is known. Through our analysis, 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 to increase market share. We hypothesize that the oldest and youngest drivers contribute the most car collisions therefore an advertising campaign to raise awareness among those groups could be the most beneficial to this motor vehicle repair business. Through this process we can recommend some actions that the company can take for incremental market share growth, and then compute the incremental volume that they can expect as a result.

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

Problem Statement

Use historic auto collision data from NYC to estimate the future collision volume. Then, take these collision estimates and combine them with known market share of the fictitious auto repair business to forecast their expected auto repair volume.

Additionally,

- Recommend a targeted marketing campaign for market share growth, and then forecast the expected incremental gains if the campaign is successful (by an assumed growth percent per demographic).
- If time permits, evaluate the effect of COVID-19 on this industry sector.

Primary Research Question

Can the historic volume of NYC Auto Collisions be used to accurately forecast auto collision volumes in a future year?

Supporting Research Questions

1. What demographic should be targeted for a business wanting to grow their motor repair market share? Put differently, do some demographics make an outsized contribution to the quantity of collisions in New York state?
2. Can we quantify the lift in revenue, and even profits, expected from such a successful marketing campaign?

Commented [CG1]: This would require us to know cost structure of running an auto shop. A bit of digging needed, probably not impossible.

Commented [CG2R1]: Found AutoZone's income statement. That could be a proxy for all auto shops (even though most of the shops we are talking about do not enjoy the economies of scale that AutoZone does) https://www.wsj.com/market-data/quotes/AZO/financials/annual/income-statement

Commented [HM3R1]: AutoZone doesn't do vehicle collision repair, although they do provide service parts.

Link to the Automotive Body and Related Repairers from the US Bureau of Labor and Statistics. Would be good if we could condense this down to a per car # at some point.

https://www.bls.gov/oes/current/oes493021.htm

Commented [CG4R1]: Good to know. Ah, the perils of not owning a car.

---

[1] We are using a fictitious auto repair shop since we have been unable to acquire the financial statements and operating records of a real auto shop given that they are privately owned businesses, for the most part, with no SEC reporting requirements.

[2] Knowing that there are 16440 motor vehicle repair shops, we assume that they enjoy perfect competition, each having about .61% of the market share. Count is accurate as of June 18th, 2022, at 1300 hours ET.

3. Time permitting, we will explore the effects of other phenomena such as
    a. Did COVID-19 change the overall auto collision volume in NYC?
    b. Insurance rates
    c. Policy and the expenditure of tax revenues: Could they be better allocated towards public transit?

Business Justification

We will use analytics, New York state collision data and our knowledge of business (marketing) to answer questions and make assertions that will allow motor repair shop owners to make decisions that will increase profitability for their businesses.

Question	Sample assertion	Why it matters
What will the demand look like in a future year, say, 2023?	We expect an average demand of 10 cars per week, each requiring about 5 person-hours	With this information, an owner of an auto repair shop is better positioned to 1. Identify opportunities for expansion. If the forecasted demand is significantly greater than the currently served demand, the owner can opt to open a second location to meet demand. The owner might even opt to rent/purchase more real estate if they determine that they will not be able to house the cars that need service. On the other hand, if predicted demand is significantly lower, the owner can then seek subletters so that their space is not wasted. 2. Meet staffing needs. If the owner learns that their shop will need to produce 50 person-hours a week when they currently are only staffed for 40, they can then seek mechanics who can help them meet demand.
Which demographics make an outsized contribution to car collisions?	Men between 20 and 24 contribute the most to the number of collisions	With this information, an owner of an auto repair shop is empowered to target the most valuable audiences with their advertising revenue.
How much additional demand, and therefore revenue, will be generated from a $100 expenditure on advertising?	A $100 expenditure on advertising is responsible for 1 extra person-hour of demand which equates to $200 in added revenue	With this information, an owner of an auto repair shop can quantify their return on as spend (ROAS) and be empowered to: 1. Justify spending their hard-earned profits on advertising knowing that they will benefit from the expense. 2. Know how much added demand their advertisements are expected to generate and therefore be ready to meet that demand.

Commented [HM5]: @Cunningham, Nicholas J @Lange, Brittany C @Chille, Lisa G @Taha, Delband I'm thinking we should put this as 2021 for the predictions. This will allow us to compute the real accuracy of our models and any recommendations? Thoughts?

Commented [CG6R5]: Ah, I used 2023 to make it clear that we are talking about the future. So basically, once all is said and done, a business owner can take our model *now* and predict demand for 2022. What do you think? Should we phrase that differently?

Commented [HM7R5]: I think we're saying the same thing. Maybe rather than listing a year we can say make predictions for the year following the data?

Commented [CG8R5]: Fixed :)

Dataset

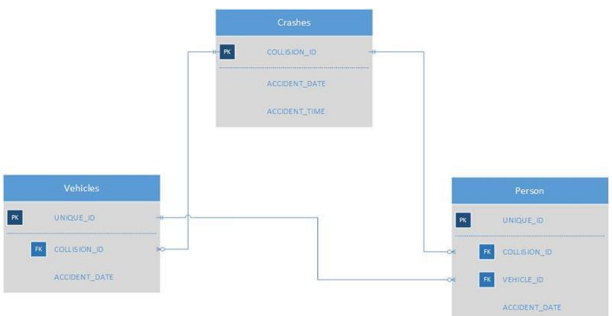
Sources

The main data sources for this project are sourced from data.gov. These data represent New York state vehicle collisions.

Name	Link	Row count	Column count
Crashes	Link	1,896,229	29
Vehicles	Link	4,692,054	21
Person	Link	3,704,406	25

Description

Below is an Entity Relationship Diagram from the NYC Dataset documentation:



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 is 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 is 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 is 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

The synthetic column indicates whether the variable will be created or if it already exists in the dataset.

Key Dependent Variable

Name	Type	Description	Use	Synthetic?
Volume of auto collisions	Integer	An aggregation of car crash volume per month	Extrapolating future demand and potential revenues.	No

Key Independent Variables

Name	Type	Description	Use	Synthetic?
CRASH_DATE	DateTime	When a crash happened	Create new time-related features	No
PERSON_AGE	Integer	A car crash participant's age	Modeling demographic factors	No
PERSON_SEX	Factor (F/M/Unknown)	A car crash participant's gender	Modeling demographic factors	No
PERSON_TYPE	Factor (Driver/Pedestrian/Passenger/Cyclist/...)	Level of participation of people in a car crash	We only care about `Drivers`. We will filter out other participants to not double count collisions.	No
TIMESTEP	Integer	Indicates the chronological position of the month at hand.	Tie all months and years together in chronological order	Yes

AGE_GROUP	Factor (15_19/20_24 /...)	A binned version of the PERSON_AGE field in 5-year increments from 15-100	Used for demographic analysis	Yes

Above is a list of the variables that we expect to be most relevant when solving the problem we have outlined.

Approach/Methodology

Planned Approach

Primary Research Question

Data wrangling

1. Join all independent datasets into a combined dataset.
2. Engineer the synthetic features as outlined above in preparation for modeling.
3. Aggregate the data to a monthly granularity.
4. Create cross-validation and test datasets. Since we are dealing with a temporal model, we will select all except the final year for cross-validation, and then use the final year for testing. Depending on the practicalities of our analysis, we might opt to further split the cross-validation dataset into training and validation datasets so we can consistently benchmark the performance of the models against each other.

Feature selection

5. Even though we have identified key predictors, we will experiment with LASSO, elastic net, and other automated feature selection methods to find whether any additional features improve the performance of the models we are training.

Training and validation

6. We will then cross-validate/train and validate a few models on the data at hand. Some models that we are considering include *simple* and *multiple linear regression*, *CART*, *boosted tree*, and *random forests*.
7. We will select a final model by comparing the R^2 or MAE of the various models at hand.

Testing

8. After selecting a final model or an ensemble of models, we will then compute the expected car crash count for the test dataset. Since we know the actual car crash volume, we will have the opportunity here to understand the accuracy of our chosen model. We will use R^2 or MAE once again to measure performance.

Prediction

9. We will then use the model to estimate demand at the local repair shop, using their market share (assumed to be ~.61%).

Secondary Research Questions

Additional project work will then focus on understanding the impact of

Commented [CJ9]: @Hedberg, Jeffrey M @Chille, Lisa G @Lange, Brittany C @Taha, Delband Im assuming for the marketing component we are just going to make hypothetical campaign assumptions on spend, lift, conversion, etc. to end up with impact. I made some wording adjustments in those areas

10. Demographics on motor vehicle collisions, therefore powering our recommendation for which demographic(s) to focus on to grow the shop's market share. Here, we expect to use multiple linear regression with squared predictors standardized around mean 0 and standard deviation 1.
11. Using simulated advertising campaign factors, including industry benchmarks for spend, conversion and lift of a successful advertising campaign to determine increase in demand at the auto shop. Apply these outputs to a costing model for an auto repair shop to derive profits and ROAS for future investments. We anticipate using a linear regression model here.

Extra

If we have time, we will also look at the

12. Impact of COVID-19 on the collision volume.
13. Whether some demographics contribute more fatalities than others, and whether those fatalities track with the number of collisions contributed by those demographics[3].
14. Impact of traffic collisions on
    a. Insurance premiums.
    b. Hospital costs, including growing medical debt.
15. Policy implications of funding public transit to reduce traffic volume.

[3] Here, we will be able to experiment with more models such as logistic regression. We will not use R^2 to estimate performance of the model when cross-validating or training, validating, and testing logistic regression models.

Anticipated Conclusions/Hypotheses

We expect that we will

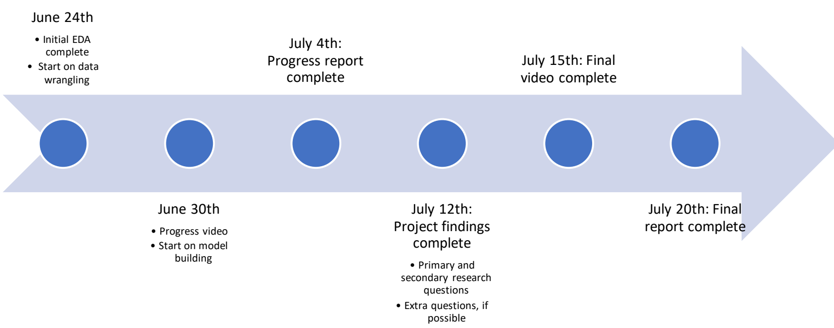
1. Be able to estimate the monthly demand for a fictitious auto shop in New York state by extrapolating the monthly car crash volume with reasonable accuracy.
2. Identify a demographic that comprises the largest market segment for motor vehicle repairs after an accident; from our preliminary investigation, we suspect that young men comprise this demographic.
3. Be able to assert that $x expenditure on advertising targeted towards the most valuable demographic will result in $x+y revenue.
4. Be able to quantify the impact of COVID-19 on demand in this sector. Our preliminary investigation has already highlighted this.

Analysis Impact on Business Decisions

As stated in the business justification section, equipped with our analysis, an owner of an auto repair shop is better positioned to

1. Identify opportunities for expansion. If the forecasted demand is significantly greater than the currently served demand, the owner can opt to open a second location to meet demand. The owner might even opt to rent/purchase more real estate if they determine that they will not be able to house the cars that need service. On the other hand, if predicted demand is significantly lower, the owner can then seek subletters so that their space is not wasted.
2. Meet staffing needs. If the owner learns that their shop will need to produce 50 person-hours a week when they currently are only staffed for 40, they can then seek mechanics who can help them meet demand.
3. Target the most valuable audiences with their advertising revenue.
4. Justify spending their hard-earned profits on advertising knowing that they will benefit from the expense.
5. Know how much added demand their advertisements are expected to generate and therefore be ready to meet that demand.

Project Timeline



Appendix

Descriptive statistics for the datasets

Crashes dataset

Crashes Dataset

```
#### 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

Person Dataset

```r
#### 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

Vehicle Dataset

```r
#### 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

New York state motor vehicle repair shop counts

By county

County	Number of auto-repair shops
Albany	327
Allegany	73
Broome	234
Bronx	638
Cattaraugus	128
Cayuga	131
Chautauqua	216
Chemung	104
Chenango	99
Clinton	143
Columbia	95
Cortland	77
Delaware	107
Dutchess	324

Erie	1054
Essex	53
Franklin	83
Fulton	106
Genesee	102
Greene	87
Hamilton	6
Herkimer	105
Jefferson	180
Kings	957
Lewis	47
Livingston	112
Madison	121
Monroe	680
Montgomery	91
Nassau	1216
Niagara	305
New York	99
Oneida	400
Onondaga	510
Ontario	162
Orange	487
Orleans	58
Oswego	193
Otsego	111
Putnam	97
Queens	1190
Rensselaer	197
Richmond	280
Rockland	256
Saratoga	214
Schenectady	197
Schoharie	63
Schuyler	47
Seneca	65
Steuben	175
St. Lawrence	191
Suffolk	1609
Sullivan	128
Tioga	71

Tompkins	95
Ulster	224
Warren	97
Washington	92
Wayne	144
Westchester	878
Wyoming	74
Yates	35

Total
16440

Average market share of each auto shop
.61%

References
Edlin, Aaron and Pinar Karaca-Mandic. "The Accident Externality from Driving." *Journal of Political Economy* 114.5 (2006): 931.