NTNU

Kunnskap for en bedre verden

TDT4259 Applied Data Science

Vehicle Collisions in New York City

Group 11

Authors:

Name	Email	StudentNr
Anhkha Nguyen Vo	akvo@stud.ntnu.no	x
Emilie Lia-Rognli	emilili@stud.ntnu.no	x
Julie Holte Motland	juliehm@stud.ntnu.no	x
Kristoffer Nyvoll	kristnyv@stud.ntnu.no	x
Magnus Lauritzen Holtet	magnulho@stud.ntnu.no	x
Sigrun Nummedal	sigrunnu@stud.ntnu.no	x
Øyvind Jalland Schjerven	oyvinjs@stud.ntnu.no	x

Table of Contents

List of Tables

List of Figures

1 Introduction and Problem Definition

1.1 Introduction

New York City (NYC) is the most populated city in the United States (US) - and it has got the traffic to prove it [1]. In the 2021 Urban Mobility Report published by Texas A&M Transportation Institute, New York was listed as the worst city for any road user in the US [2]. Furthermore, according to the World Health Organization, approximately 1.3 million people die each year in motor vehicle collisions, and 3% of all gross domestic product worldwide is spent as a result of road traffic collisions [3]. In other words, there are clear sociopolitical and economic incentives to reducing motor vehicle collisions in general, and New York City has many areas of improvement.

From 2016 to 2020, there was a total of 1 016 774 reported motor vehicle collisions in New York City. Measures have been attempted to reduce the number of crashes, but none of these measures have prevailed in recent years. Aside from the obvious consequence of death and injury, unsafe traffic systems entail enormous costs for repair, court proceedings, traffic jams may even invoke public fear of using the roads. The data set containing details about every single motor vehicle collision in NYC might hold the answers to fixing the problem.

Nearly every inhabitant in NYC relies on transport in some way. Therefore, motor vehicle collisions can have severe ramifications for a vast majority of the inhabitants. Providing a safe system for pedestrians, cyclists, and drivers is consequently essential for developing any modern city - and all the more important for cities on the scale of New York City.

1.2 Problem Definition

This report is based on the data set "Motor Vehicle Collisions - Crashes" from New York City open data [4]. This report aims to suggest clear and feasible recommendations and focus areas on reducing accidents and improving traffic safety in New York, captured from insights gained from the data analysis. These recommendations were discovered by separating the available data into different categories and analyzing how factors such as locations, causes, and time of day influence the accident statistics. This knowledge is helpful in several ways, and the New York Police Department (NYPD) can for instance increase preparedness in the areas and times that are most exposed. Moreover, the most problematic intersections and locations can be uncovered, aiding the New York City Department of Transportation (NYC DOT) in making NYC a safer place for all road users.

The following questions can be answered by analyzing the data set:

- What are the most common causes of motor vehicle collisions?
- Where do vehicle collisions most commonly occur?
- When do vehicle collisions most commonly occur?
- Which types of road users are most involved in vehicle collisions?

1.3 The Team

Seven 4th year informatics-students wrote this report. All members knew each other beforehand and had previous experience working together in school projects and through work. This familiarity made cooperation effortless, and there was no need for a group leader. The flat structure implied that all responsibilities were assigned according to the work tasks and interests of the group members. Workload distribution was even throughout the entire project period, which lasted for a little over two months. The team decided collectively on the data set and the other details about the project. Most of the data analysis was handed out by Sigrun, Emilie, Julie, and Anhkha, while Kristoffer, Øyvind, and Magnus took overall responsibility for different parts of the report. Everyone contributed to every part; the "responsible group member" just had the overall responsibility to ensure progress and quality.

All group members met for a physical work session twice a week to facilitate optimal collaboration and efficiency. This allowed for democratic and agile working methodology since all questions and remarks could be collectively answered and dealt with when they arose. In retrospect, meeting up twice a week throughout the project was a huge success, ensuring an early start and better knowledge sharing internally.

2 Background

Data analytics as a tool for decision making has almost become a necessity for handling complex business objectives. With the current rate and complexity of data production of most companies, manual analysis of the data is no longer a sustainable option [5]. Yet, it is crucial for companies to take advantage of the available data in the biggest extent possible. This combination propagates the validity of data science as a tool for analysis and decision-making.

In the private sector, there is mainly a monetary incentive to utilize data science. In the public sector, however, this incentive is not as predominant, but still gaining more traction as the benefits of data analytics is becoming wide-spread. In both sectors, there is therefore potential for improved use of available resources as a result of the measures that can be discovered through data science. Exploiting this potential to generate public value is now becoming more widespread in the public sector [6].

2.1 Definition of Objective

In the early 1990s, NYPD was struggling with cultural issues, limited communication and lack of up-to-date crime statistics. As a countermeasure, a program called CompStat, short for computer comparison statistics, was started. CompStat used crime statistics to generate greater awareness and control of the immediate on-the-ground crime situation. Following the introduction of CompStat, there has been a dramatic drop in crime rates in New York City. This drop is largely a result of the increased awareness and flow of information caused by the introduction of CompStat, as well as other measures implemented based on the gained statistics [7].

Between 1993 and 1998, homicides dropped 67 %, burglary was down 53 %, and robberies were down 54 % [8]. This success led to the NYPD examining other areas for implementing data-driven policing. The TrafficStat program was started in April 1998 due to the success of the CompStat program. The overall goal was to improve traffic safety in New York City through data-driven decisions. With TrafficStat, the department tried to pinpoint the worst locations and deploy additional officers to these locations to watch for dangerous, illegal, or drunken drivers. In addition, the department desired to analyze which primary factors contribute to traffic accidents at these locations and what NYPD could do to implement countermeasures. Unlike the CompStat program, which has dramatically reduced crime rates after its implementation, TrafficStat does not seem to have had a significant effect on motor vehicle collisions to this point. On the contrary, there was an increase of approximately 12% in total collisions in New York from 2014 to 2017 [9].

2.2 Decrease Collisions in New York City

Using the data in this data set, however, might provide the necessary insight to actually reduce the number of collisions. Data has one particular advantage over the human mind; it does not tell lies, and is largely unaffected by our biases. Collected by police officers at collision sites, the data in this data set should provide a complete overview of the collisions in NYC. When introducing new measures to improve the situation, insight from this data can help the NYC DOT make educated decisions without the biases that previously prevented them from making a real impact. Evidence of this can be seen by looking at the results from 2014 to 2017, in which the number of collisions increased despite their efforts to reduce them [9].

The overall objective of this project is to decrease collisions in NYC. Considering the complexity and size of this task, it is preferable to narrow down the scope into two smaller objectives. These will be central to the success of the overall objective. The first objective is to decrease the number of collisions on the most exposed roads. The second objective is decrease the number of collisions based on the most common factors. Both of these two objectives will be the focus in this report.

2.2.1 Decrease the number of collisions on the most exposed roads

The simplest way to mitigate a significant amount of the collisions is to identify the most exposed roads and intersections, and make improvements on these. The data on the most exposed roads is quite clear - the total amount of collisions on for instance Belt Parkway and Broadway from 2016 to 2020 is close to 18 000. This will be discussed in greater detail later in the report. Improving the traffic management at these locations would mitigate a large portion of these collisions, thereby reducing the total number of collisions effectively. By directing focus to the biggest "contributors", NYC DOT have the highest return for their efforts.

2.2.2 Decrease the number of collisions based on the most common factors

Examining the most common factors of collisions will provide NYC DOT with a helpful overview and is an important part of minimizing the amount of collisions. Understanding why crashes occur will assist the NYC DOT in determining where they should focus their efforts in order to achieve the overall objective.

Sadly, people will always be prone to making mistakes no matter how well the roads are designed, and as section 4 will elaborate further on, many of the common factors can be labeled as "personal errors". Unlike finding the most exposed roads and improving them, altering the behaviour of road users is a lot less straightforward. Luckily, large-scale behaviour-change campaigns are indeed feasible to implement, and have been shown to work well in the past. In Norway, for example, not using the seat belt was found to be a contributing factor to the increased collision mortality. A campaign was initiated with the purpose of making more people use the seat belt, and in 2019 97,4 % of drivers and passengers used the seat belt [10].

2.3 Data Strategy

The selected data strategy was CRISP-DM, which stands for "Cross-Industry Standard Process for Data Mining." CRISP-DM was compared to BAM (The Business Analytics Methodology) a different data method. CRISP-DM is a cross-industry standard that can be used in any data science project, regardless of the domain [11]. This also makes the strategy well-documented which was beneficial for the group since everyone was new to data science. Compared to CRISP-DM, BAM has less documentation, most likely due to lesser popularity. BAM strategy is also more targeted towards a business domain which is not that applicable to New York as a city and it's residents. BAM could be considered more as a tool that provides a logical structure and logical precedence of activities that can be used to guide the practice of analytics (i.e., a mental model), compared to CRISP-DM which provides a process that can be followed systematically [12]. CRISP DM's documentation and well-defined procedure may have drawbacks, such as being excessively rigid and denying the group freedom due to strict documentation [13]. Despite this, the group chose to use CRISP-DM and adjusted the approach to the group's advantage.

2.4 Approach with CRISP-DM

CRISP-DM is a process of six phases, and includes virtually everything necessary to carry out a successful data science project [14, 15]. All the phases are further elaborated upon in subsections below. Together, the insights they provide will aid the NYC DOT and their TrafficStat program. In short, TrafficStat aims to decrease the number of collisions. Traffic analyses that previously did not use data science to validate their hypotheses can use this insight to test the validity of these and discover new information.

2.4.1 Business Understanding

The first step is to gain an understanding of what sort of goals would be beneficial for the organisation to focus on. In order to facilitate for insightful data analysis, the establishment of business objectives are necessary. However, such objectives does not always correspond to a data problem. Therefore, these objectives must be transformed into result based problems with connected data sources to constitute a better basis for the analysis. Section 2.2 explains the strategy for achieving the business goals.

2.4.2 Data Understanding

Assuming that the necessary data has already been collected, one should start with the initial screening of this data. With the unprocessed data collection at hand, the first obvious step is to get familiar with the meta data, data fields and explore some initial hypotheses. In this phase it is also important to identify quality problems with the data and gather first insights into patterns and interesting subsets [15]. All the limitations and weaknesses are discussed in section 5.

Discovering at an early stage if the data set lacks critical data or whether the data is misleading or incomplete is absolutely crucial. Investing substantial time working with a data set only to later find that any results will be invalid due to problems with the data set implies a lot of wasted time. This can easily be avoided by following this step.

2.4.3 Data Preparation

Effective data analysis requires that the data is properly prepared. Most raw data sets usually need extensive cleaning before the actual analysis can begin. In the data preparation phase, activities to construct the final data-set is performed [15]. In order for data to be analyzed, it is important that everything corresponds to given formats and conventions. To achieve this, one first has to determine what is relevant data and exclude the irrelevant. The selected data then has to be cleaned in order to ensure high quality and correctness. Typical cleaning techniques is removing duplicates, fixing structural errors and filtering unwanted outliers. One might also need to infer or fill in missing data.

A common practice is to aggregate existing data into more relevant data, for instance calculating Body Mass Index (BMI) from height and weight. The last action in this step is to format the data so it is ready to be modelled.

2.4.4 Modelling

In the modelling phase, one evaluates, selects and applies modelling techniques fitting to the case. Choice of modelling technique for each model should be based on the available data, and the sort of information intended to be visualized. In this step, it is common to revisit the data preparation step if errors or limitations to the data are discovered. The main objective in this phase is to generate models that are useful to analyze in the evaluation phase.

2.4.5 Evaluation

In the evaluation phase, performance of the models that were created in the previous phase are evaluated. The performance of a model is based on how well the model is answering one or more of the business objectives. Consequently, the business objectives should be continually revisited during this phase. A key factor in data science is to ensure that the analysis is of value to the business objectives. If this is not the case, the whole analysis can end up being useless to the organization. Failing to be of use to the initial purpose is a common problem in data research. The findings should therefore be inspected, and the process as a whole should be examined in addition to the model evaluation.

2.4.6 Deployment

The last phase in the CRISP-DM approach, deployment, has two main objectives. The first objective is to interpret the results of the analysis, and furthermore translate these interpretations into actionable business recommendations. The second objective of the deployment phase is to secure the arrival of new data. Continuous collection of data is a necessity to enable further analysis, and thus maintain long term value. However, this is not an issue in this case, as the collection of data is already handled by the NYC DOT and the NYPD.

2.5 Synergy with Design Thinking

CRISP-DM is a great data strategy, but when used alone it might make people susceptible to "tunnel-vision" too much on the data-set and therefore forgetting about the bigger picture. This is where design thinking methodologies can be of great value. Design thinking strategies like the Double Diamond framework ensures that the project delivers something of value to the "customer", and is helpful for defining exactly what challenges should be addressed [16]. Double diamond alternates between widening and narrowing the scope, through creative/exploratory thinking and focused/analytical thinking. Figure 1 illustrates this process.

This framework has four steps: **Discover, Define, Develop and Deliver**. The three first were used together with CRISP-DM. Discovering what challenges to focus on, and thereafter precisely defining them, is essential to perform an effective data analysis. These first two steps from the Double Diamond framework were used in the Data Understanding-step of CRISP-DM (See section 2.4.2. This helped the group brainstorm different hypotheses, challenges, solutions and possible insights. Following this exploration of different opportunities, the define-step helped formulating the objectives in section 2.2.

Finally, the develop-step came in handy as part of the Modelling-step in section 2.4.4 and when transforming all the insights gained in the analysis-section into the recommendations. Overall, design thinking used together with CRISP-DM proved to be a useful combination.

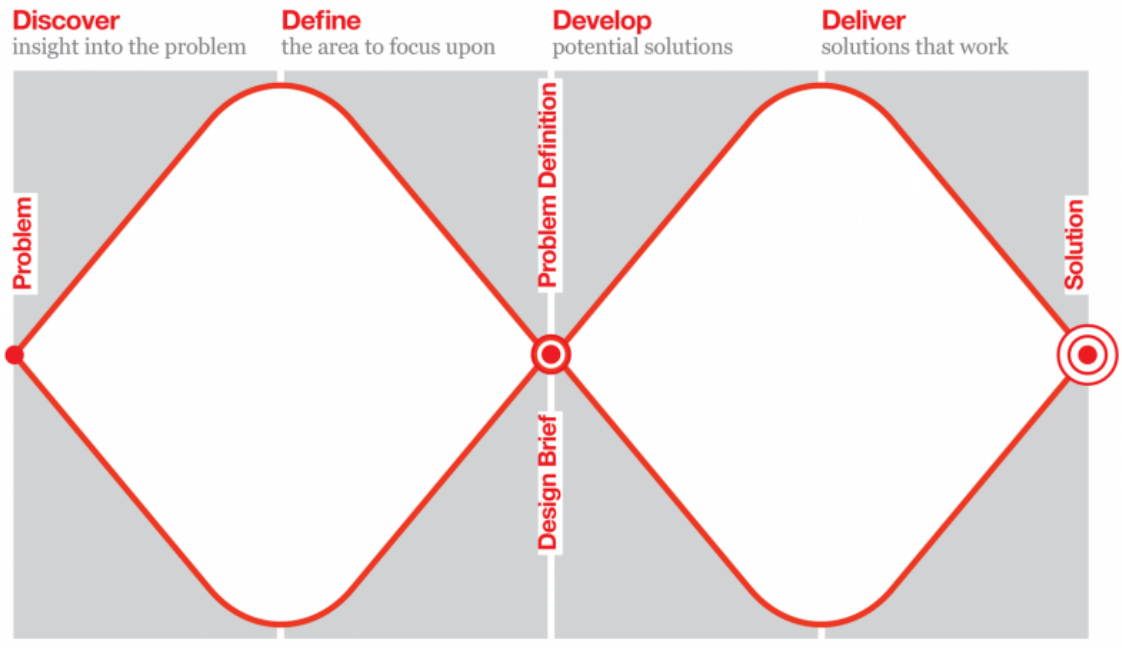


Figure 1: The Double Diamond design thinking framework.

3 Method

3.1 Description of Data Set

According to NYC OpenData, the data set 'Motor Vehicle Collisions - Crashes' contains details on all police-reported motor vehicle collisions in New York City [4]. The police files a mandatory report in cases where someone is injured or killed in a collision or $1000 worth of damage has been inflicted. Every collision is reported by the police and and stored in the data set.

The data set contains 1.82 million rows of such reports as of the 4th of October 2021. Although the data dates back to 2012, this research will only use a portion of it, from 2016 to 2020. The reasons for this decision are that (1) the reporting conventions have changed, leading to inconsistencies within the data set, (2) new roads have been built and others have been reconstructed and (3) that the overall traffic situation has changed. In addition, the most accurate picture of the current situation will be based on data from the previous five years. Given the aforementioned reasons and that there are a sufficient amount of data in the period 2016-2020, the group decided to delimit the data set to this period. Data from 2021 was removed since the is not yet completed and the data is continuously changing.

3.1.1 Attributes and Features

Each row in 'Motor Vehicle Collisions - Crashes' is divided into 29 unique columns. Many columns represent the same information using different data types, but only some of the columns are often required to bring forth the information. For instance, the data set splits the location data into eight columns. Data points such as longitude and latitude can provide a detailed description of the location. However, the street names may also provide relevant information that longitude and latitude do not. As a result, both of these location descriptions can serve as a natural point of reference in different parts of the analysis provided in this report.

The different data can be used to answer the question identified in section 1.2. The most critical data is the number of people killed/injured by the collisions, as the aim is to decrease the number of collisions and the number of people killed/injured. Those columns are also explicitly given for pedestrians, cyclists, and motorists. Also, the crash time and date can give valuable information to forecast trends.

The location data can be used to find out where most collisions occur, using both longitude and latitude as well as street name. Contributing factors of the collisions are also provided in the data set, this data is also interesting and important to look at to understand how to reduce the number of collisions.

3.1.2 Quality of the Data

The data is not necessarily the most accurate, as it may change in cases of reports being amended based on revised crash details. Further, the quality of the reports varies. Human errors and varying levels of detail in the reports can significantly influence the quality of reports. Typographical errors result in deviation in the different columns. For example, we can look at the contributing factor 'illness.' Thousands of reports have this contributing factor written with a spelling error, resulting in misrepresenting the total distribution of causes.

Moreover, more than a third of the data set has "unspecified" listed as the reason for collisions. Having such a significant amount of anonymous data severely limits the number of collisions it is possible to analyze. This limitation will make answering some of the problems more difficult, as the analysis will not cover all relevant events. There is a risk of displacements in the statistics caused by many unspecified collisions, as many of these collisions possibly can have the exact cause. If this is the case, potentially important information will not be discovered in the data analysis. The category "collision factors" is, therefore, somewhat prone to the Survivorship bias [17].

In section 5 the data quality and limitations are further discussed.

Despite these factors, the integrity and quality of the data set are found to be more than sufficient to highlight important tendencies and factors of importance given the tools provided. This integrity and data quality can be substantiated by the fact that NYC already uses the data for analysis for their TrafficStat program [4].

3.2 Methods and Tools

3.2.1 Analysis

During the data preparation, several processes have been implemented to analyze the data. The first process was to go through the data set and filter out invalid data. This data included rows with zero-values, insufficient information reported, unspecified factors for the crash, etc. This process was done first to clean up the data set before implementing other processes.

A process that was important for several parts of the data set was the grouping of equivalent values. As mentioned in section 3.1.1, there are several man-made errors in the reports. Most of these were spelling errors or different ways to spell the same information, i.e. shortened street names. Many could be fixed by grouping reports with typos together with their correctly spelled equivalent. This correction was done for, among other things, street names and common factors for collisions.

Aggregation of data is another helpful process that was implemented. Aggregation is the process where two or more types of data are combined to generate a new data type. In this report, aggregation was used to generate statistics of deaths and injuries per collision. A specific example of how aggregation was used can for instance be that latitude and longitude was combined to pinpoint the exact location of crashes. This was then used in the visualizations like Heat Maps (See figure 6).

3.2.2 Visualization

When creating visualizations, it is of utmost importance to consider how the human mind processes information. For short-term memory, which is used when analyzing graphs and other visual components, there are limitations to how many things can be stored in the mind. Visualization theory, therefore, recommends limiting the number of lines and bars in plots [18]. Considering this, the visualizations in this report are designed to be as easy to understand as possible. To achieve this, graphs are limited to only display one category of information. For line graphs, this means that only one type of line is displayed for each graph. For bar graphs, several bars are displayed but with the intent to display general tendencies.

Another factor to consider is color selection. Although complete color blindness is rare, there are many color deficiencies in the population [19]. To counter this, colors that are easily distinguishable for people with color blindness have been used. These color schemes are fetched from Color Brewer 2 [20], as recommended in the lectures.

PowerBI

Visualization of data is crucial when operating with big data sets to get an overview and gain insightful information about the data provided in the data set. The group decided to use Power BI to visualize some of the data. The tool has a simple user interface for creating different types of graphs and diagrams to represent the data provided. However, PowerBI is not supported on MacOS or Linux, which is the operating system most of the group members use. Therefore, a lot of the graphs are generated in Tablaeu.

Tablaeu

Although PowerBI is a powerful tool, the group had trouble visualizing the heatmap based on the concentration of collisions. In Power BI, the heatmap was red, alas hot, on almost all spots in New York City. This problem did not arise with Tablaeu, where the construction of the heatmap was easy and user-friendly. In addition, Tableau provides a user-friendly and intuitive user interface. Since most of the team members used MacOS and Linux, Tablaeu was used for the majority of the illustrations.

3.3 Pre-processing of Data

Conveniently, all the necessary data had already been collected and published to the NYC Open Data website. Thus, the group could start the initial screening. Some of the initial hypotheses that had to be verified include the following:

- There are a significant amount of annual collisions.
- There is a problem that needs to be fixed.
- The data is detailed enough that insight can be extracted.

If true, these hypotheses prove a need for change, and using data science to provide necessary insight into this problem is viable. Following the "Data Understanding"-step described in section 2.4.2, the data was explored, and the hypotheses were verified. The built-in analysis tool on the NYC Open Data website was used to get the initial overview over the data set.

Another important part of this initial exploration, was checking the amount and quality of data, in addition to the time period over which the data was collected. The meta-data that had to be checked out were start and end dates of the set. Furthermore, it was necessary to assess the validity and relevance of all the available columns. To increase the quality, all NULL values and unspecified values were removed from the data set. Typographical errors and varied spellings of the same information were a concern, as mentioned in section 3.2. This was handled by grouping them together, for example "Belt Parkway" and "Belt Pkw" was combined into one group.

After this initial cleaning, no major flaws or deficiencies could be found in the data points, and also the duration was sufficiently long for the data to be valid.

4 Analysis

This section will present the analysis of collisions in New York City for each of the two objectives. However, first, a general analysis that applies to both of the objectives will follow. The analysis is based on the CRISP-DM methodology described in section 3.

4.1 General Analysis

There are some essential explorations from the data that are relevant for both objectives. The first is to see the overall amount of collisions over the last five years, presented in figure 2. The overall amount of collisions are constant from 2016 to 2018. From 2018 there has been a reduction in collisions, and a drastic reduction occurred in 2020. It is natural to believe that the COVID-19 pandemic is to blame for the drop in collisions in 2020. Fewer individuals were out in the traffic because it was advised to stay inside.

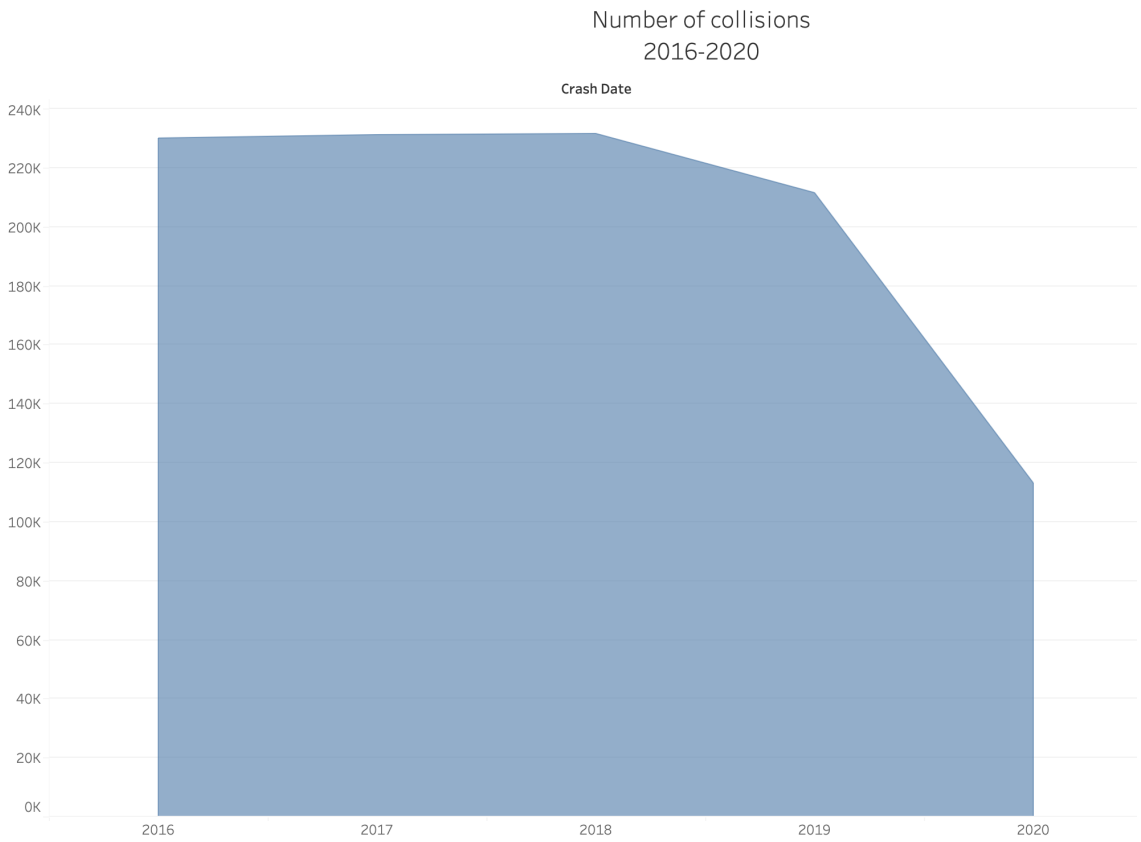


Figure 2: Monthly collisions from 2016 to 2020.

Another observation is that the collisions vary notably based on the time of the day. As figure 3 visualizes, one can see that most collisions occur from 14:00 to 18:00. It is reasonable to presume this is due to rush hour; however, there are still more collisions during that period than during the morning rush. This may imply that drivers are less observant in the afternoon than they are in the morning. The area chart also displays an even level of collisions from around 19.00 until 00.00, followed by a drastic decrease in collisions from midnight until the morning traffic.

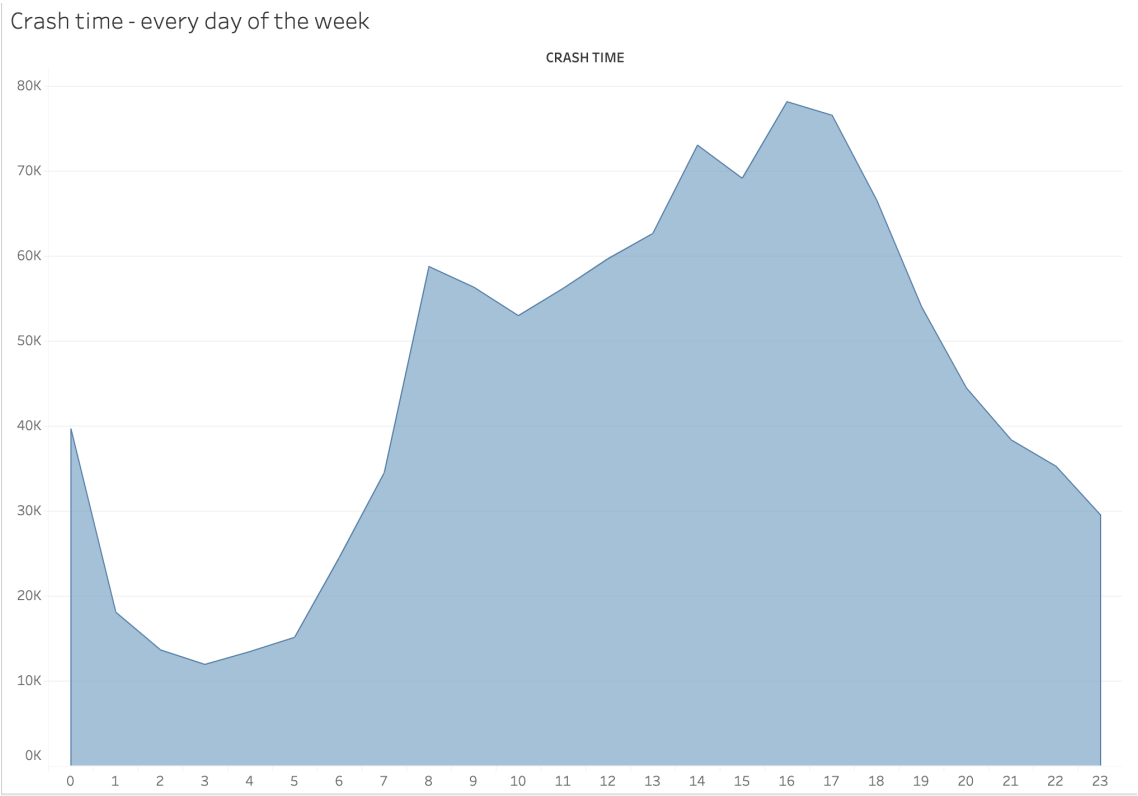


Figure 3: Time of day when collisions happens the most

Interestingly, more collisions occur at night on Saturdays compared to the overall average, as shown in figure 5a. Comparing this to Mondays emphasises how problematic this increase is, as seen in figure 5b. This is probably because more people are going out on Saturday nights than on any other weekday. One final observation is that most of the collisions occur during the weekdays, probably due to more traffic in general. This is shown in figure 4.

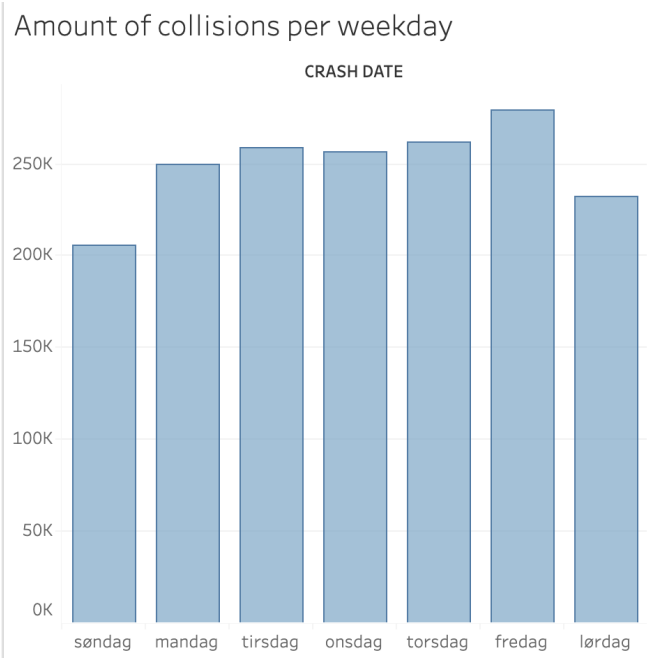
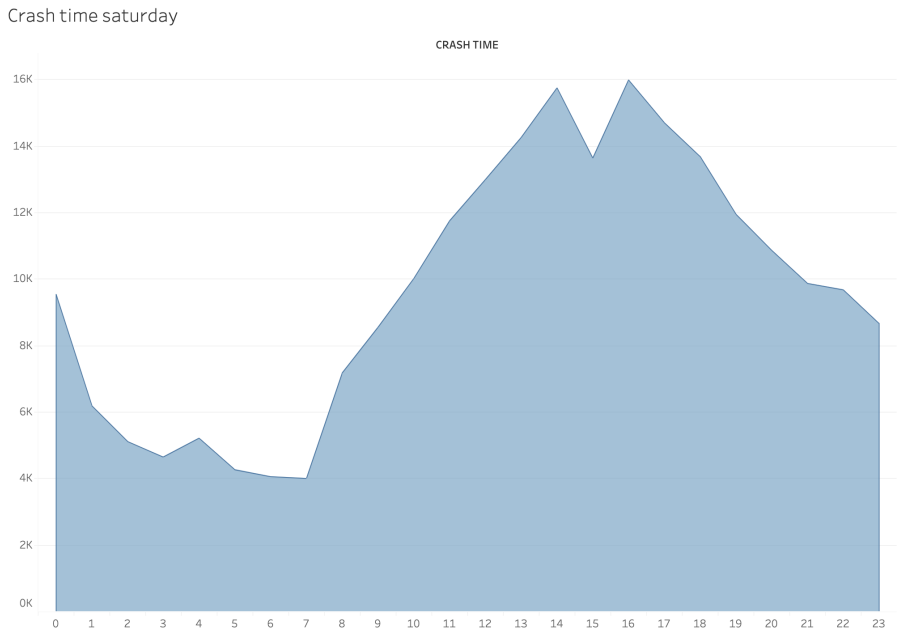
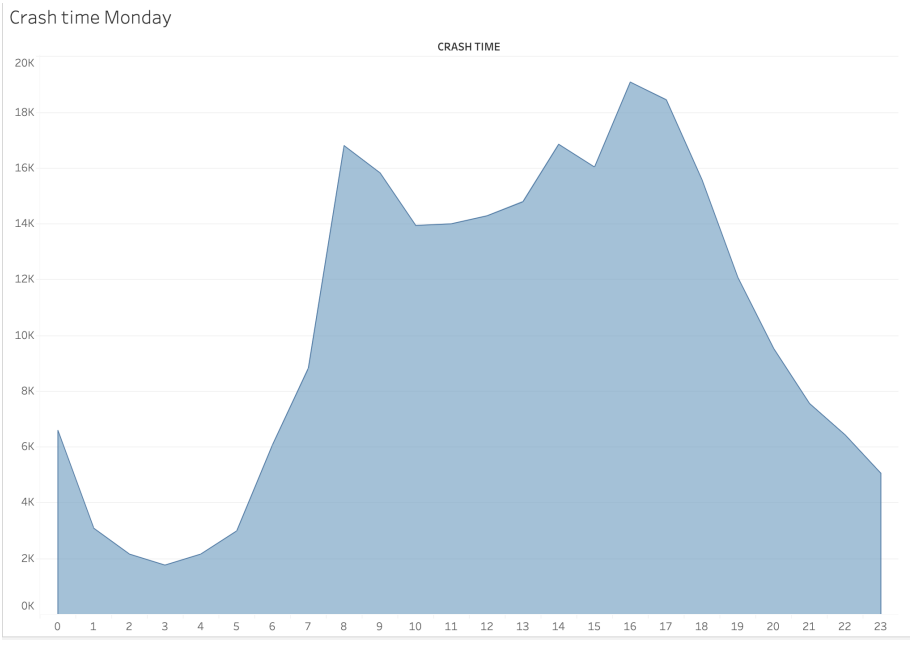


Figure 4: Number of collisions per weekday



(a) Saturday



(b) Monday

Figure 5: Collisions on specific days

In figure 6, a heatmap is presented with an overview of where collisions occur in New York City. Many collisions emerge in the center of New York City, namely in Manhattan, Brooklyn, and Queens. These places have a lot of inhabitants and workplaces, and many people bypass the areas to go to other districts. This results in much traffic, which is probably one of the reasons why the number of collisions is high at these places.

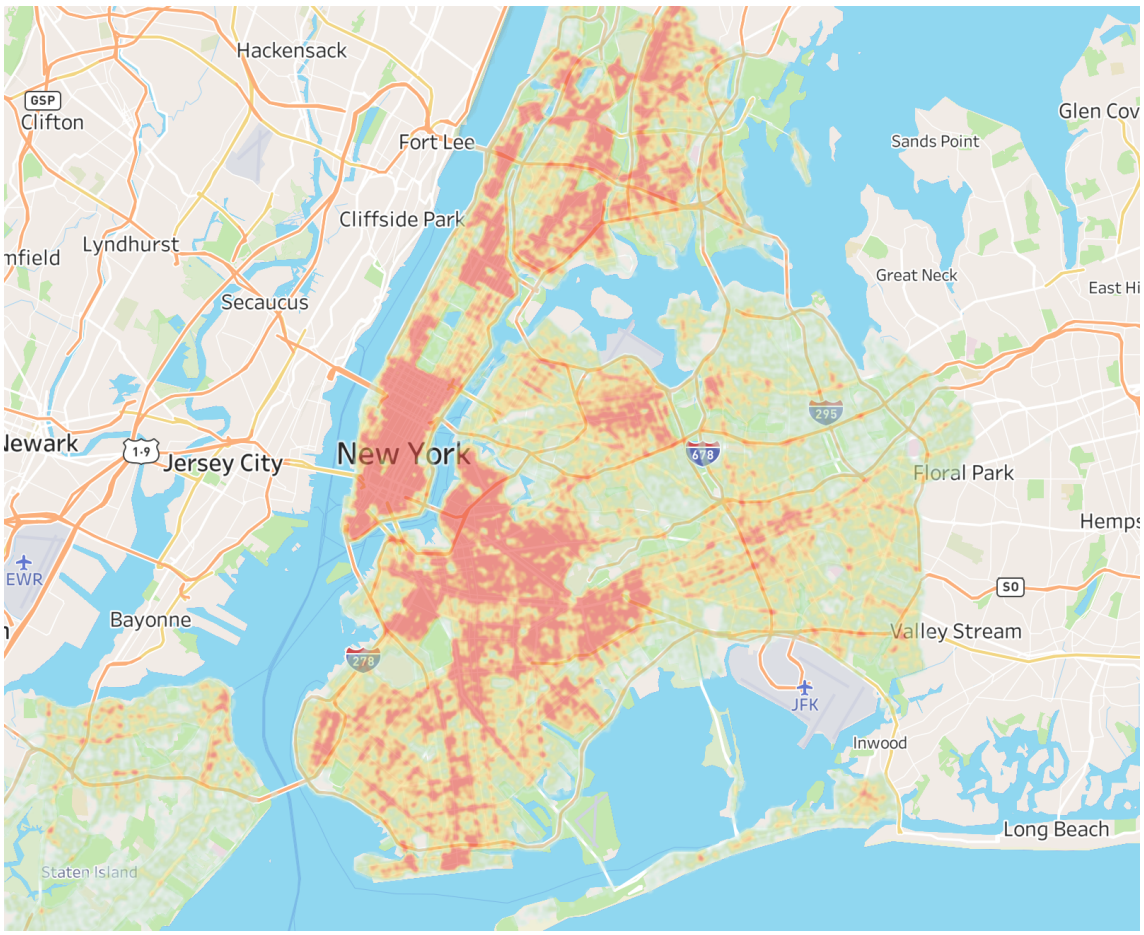


Figure 6: Heatmap over collisions in New York City (2016-2020)

Based on the overall observations, the data will be further analyzed to understand where and why collisions occur.

4.2 Decreasing number of collisions on the most exposed roads

To accomplish the objective of decreasing the number of collisions on the most exposed roads, it is first necessary to analyze where all the collisions take place. As seen in Figure 7, the most exposed road for collisions is Belt Parkway. Belt Parkway has a length of approximately 24 miles (41 km) and runs through Brooklyn and southern Queens [21]. The top four roads shown in figure 7 are highways, meaning they are major, well-constructed roads meant for travel and carrying heavy traffic. It is not until fifth place that a non-highway road is the most exposed for collisions. Broadway is one of the most popular tourist attractions in New York City, and in season 2018-2019, it sold for more than $1.8 billion in tickets, implying that many people visit and travel Broadway [22]. Belt Parkway and Broadway will be the roads of interest in this objective since there are two different types of roads and may therefore have different causes and measures. Together, these two roads were deemed representative for the problem.

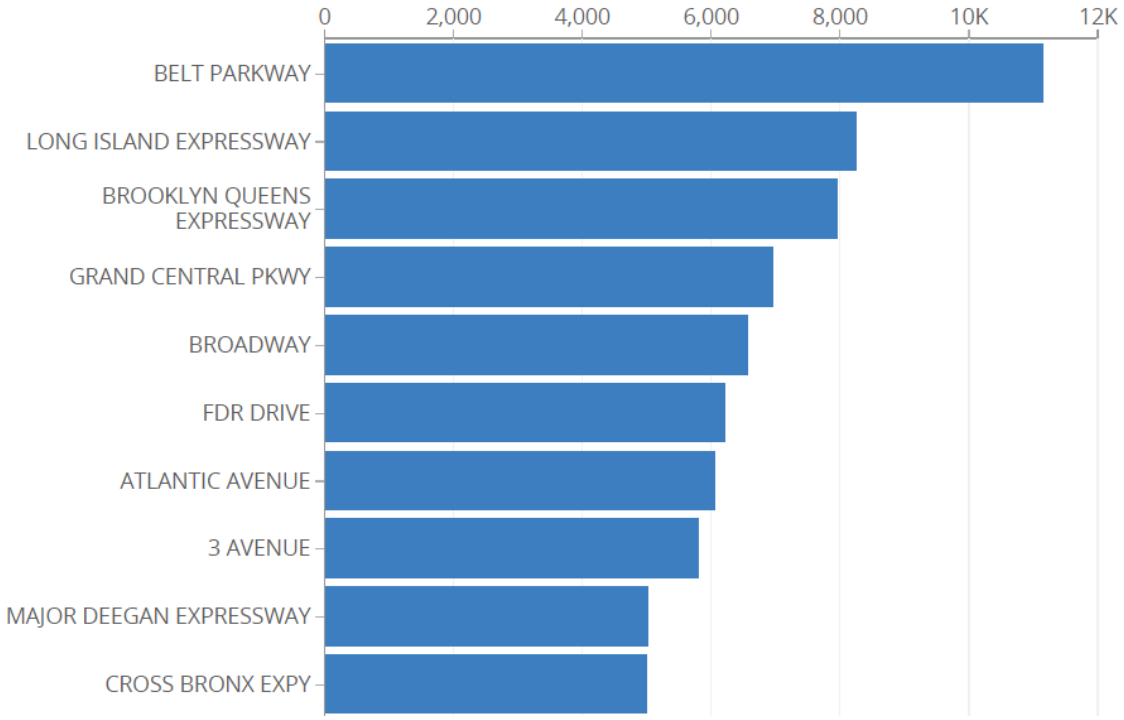


Figure 7: Top 10 street addresses with most collisions from 2016 to 2020.

The next step was to analyze these two roads individually to see if there existed any characteristics in why the collisions occurred, the severity of the collisions, and the death and injury rate.

4.2.1 Why Do Collisions Occur?

Knowing that Belt Parkway and Broadway are two roads with many collisions, it was necessary to examine why the collisions occur. The contributing factors for Belt Parkway can be seen in figure 8. The most contributing factor is *driver inattention and distractions*, including different elements like using the phone, eating, distractions outside, and so on. This will be further elaborated in the analysis of the other objectives as drivers' inattention is the number one contributing factor in general. Therefore, other factors will be instead analyzed.

The NYC DOT should also be concerned about factors that relate to the roads and the traffic. *Following too closely* is the second factor and accounts for nearly the same amount of collisions as *driver inattention*. This could indicate that there is too much traffic on the road. There could be a lot of traffic queues and various speeding resulting in vehicles following too closely. Belt Parkway is a highway with lots of heavy vehicles, queues, and various speeding will make it challenging to keep a steady pace. Queues and congestion would be two aspects to look at for the NYC DOT.

The following two most common factors for collisions are *unsafe speed* and *unsafe lane changing*. *Unsafe speed* probably follows from the fact that the speed limit at Belt Parkway is high. *Unsafe lane changing* could also be a result of an aspect discussed above - simply that there is too much traffic on the roads. Logically, increased number of vehicles on the road will probably make it more unsafe, and therefore more dangerous to change lanes.

The most common factors for collisions on Broadway can be seen in figure 9. Similar to Belt Parkway, *driver inattention and distractions* are the number one contributing factor. However, the second factor is *failure to yield the right-of-way*. There could be several reasons for this. One of the reasons could be distracted drivers, failure to adhere to the rules, or simply not knowing the rules about yielding for other vehicles. Lastly, the view could be limited or obscured.

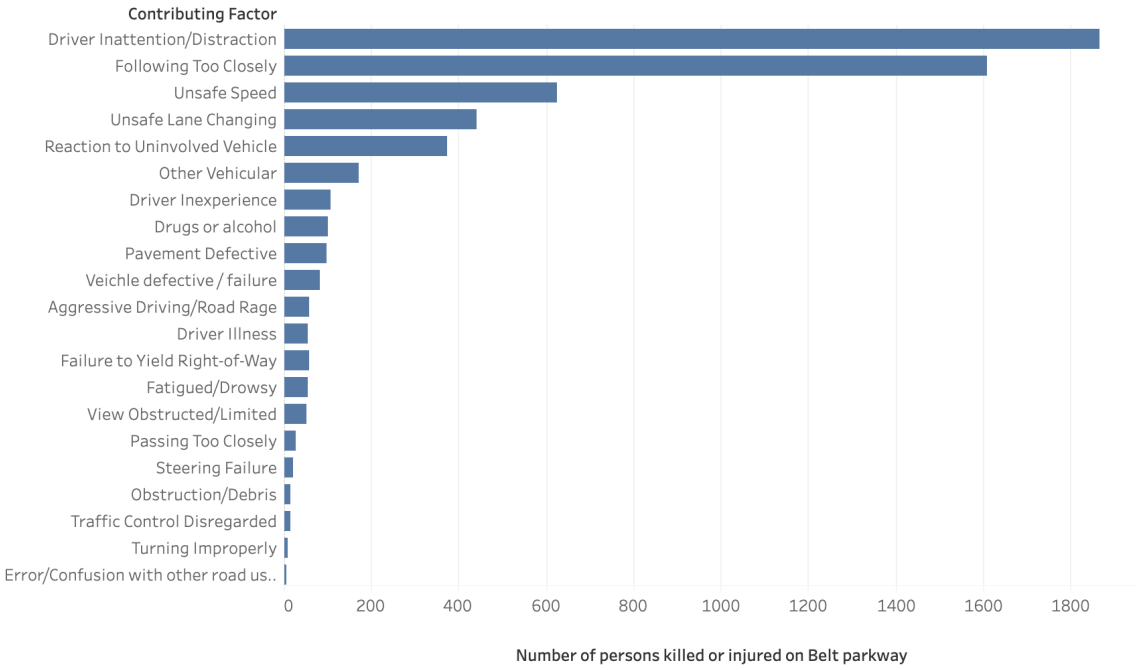


Figure 8: Contributing factors for collisions on Belt Parkway

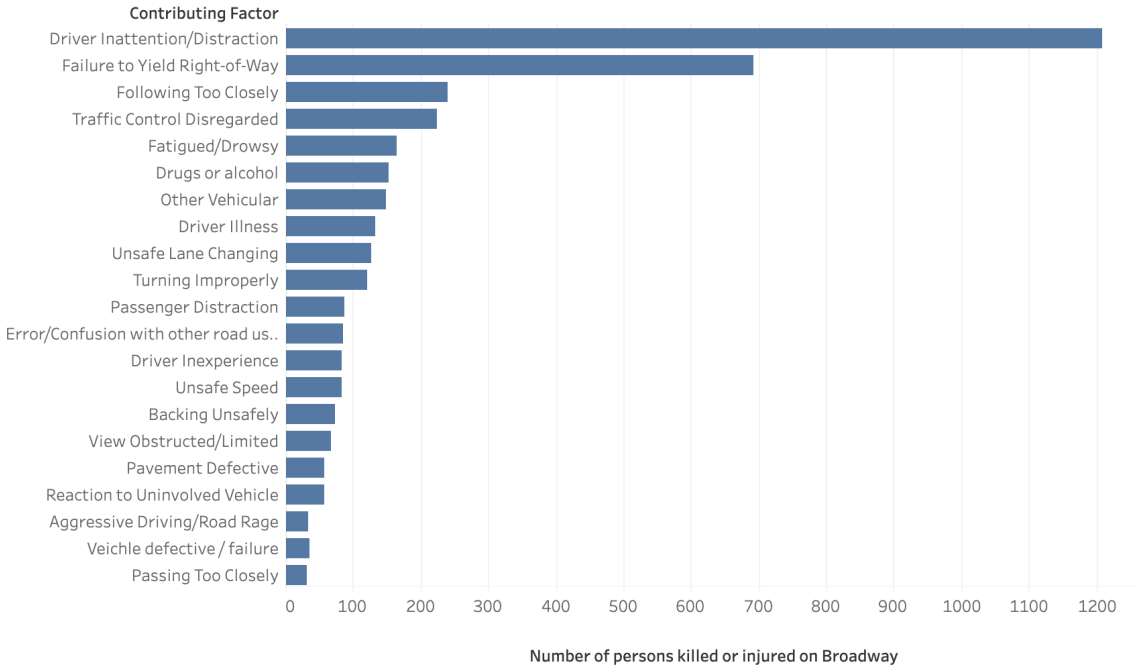


Figure 9: Contributing factors for collisions on Broadway

The following factors are *following too closely*, and *traffic control disregarded*. Broadway has more traffic lights and signs than Belt Parkway, possibly resulting in more traffic violations. This follows naturally from the fact that there are more traffic controls that can be disregarded. *Failure to yield right-of-way* can also be below this factor, as that is a disregard of traffic control. Also, *drugs and alcohol* are highly rated factors. This may be due to the driver or pedestrian being under the influence of alcohol or drugs. Broadway is a street with pedestrian fields, crossings, bars, restaurants, tourism, and lots of people, resulting in an increased risk of hitting a drunk person with a vehicle. Alcohol makes people inattentive, which may lead to people crossing streets without checking.

Considering all this, NYC DOT has a solid foundation for why collisions occur, which should be used actively to prevent collisions. Looking at the contributing factors, NYC DOT knows what causes the most collisions and should accordingly act upon it.

4.2.2 Severity of the Collisions

Figure 10 and figure 11 are created with Tableau to present the severity of the collisions. The figures represent how many injured and killed motorists, pedestrians, and cyclists are for Belt Parkway and Broadway.

In figure 10a there are mostly motorists that are injured, and almost no pedestrians or cyclists were killed. There is a reasonable explanation behind this as Belt Parkway is a highway meaning few to no pedestrians and cyclists are allowed on the road. In 2018, closely followed by 2019, the most injured motorists on Belt Parkway with over 1200 per year. Compared to figure 10b there are a lot more injured people than killed. Even though the most injured people in 2018, this was the year the least people were killed. 2019 and 2020 had the most skilled people with 6 persons per year, of which 4 were motorists, and 2 were pedestrians. This indicates to which road user and how NYC DOT should distribute their resources.

Belt Parkway is a long highway making it more likely to have several collisions. Belt Parkway also serves as the main route to JFK International Airport and commuter route for workers and travelers. Therefore, the highway is often subject to a high level of congestion. Because there are a lot of driving vehicles, the collisions may often involve several vehicles. On September 20th in 2021, at least 15 people were hospitalized after a pileup on Belt Parkway. Up to 10 vehicles may have been involved in the crash [23], and this may be an indication that each collision on Belt Parkway involves many people and vehicles.

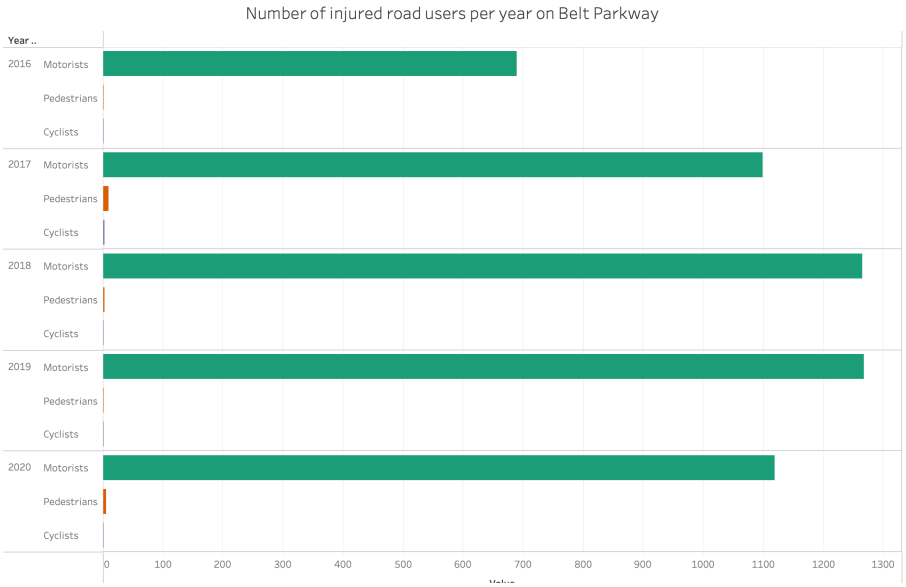
Figure 11 presents a number of injured and killed people on Broadway. This chart is very different from Belt Parkway as more cyclists and pedestrians are involved in the collisions due to Broadway not being a highway.

In figure 11a the number of injured motorists is decreasing and is at its lowest in 2020. The total number of injuries is also lowest in 2020. This could be due to the global pandemic COVID-19, where people stayed more at home than in previous years. However, in the time period 2016 to 2020 the number of injured cyclists is the highest in 2020. A reason for this could be that more people cycled to their job instead of taking public transport. Hence, more collisions involved cyclists. The number of injured people from 2016 to 2020 is overall lower than injured people on Belt Parkway in the same period.

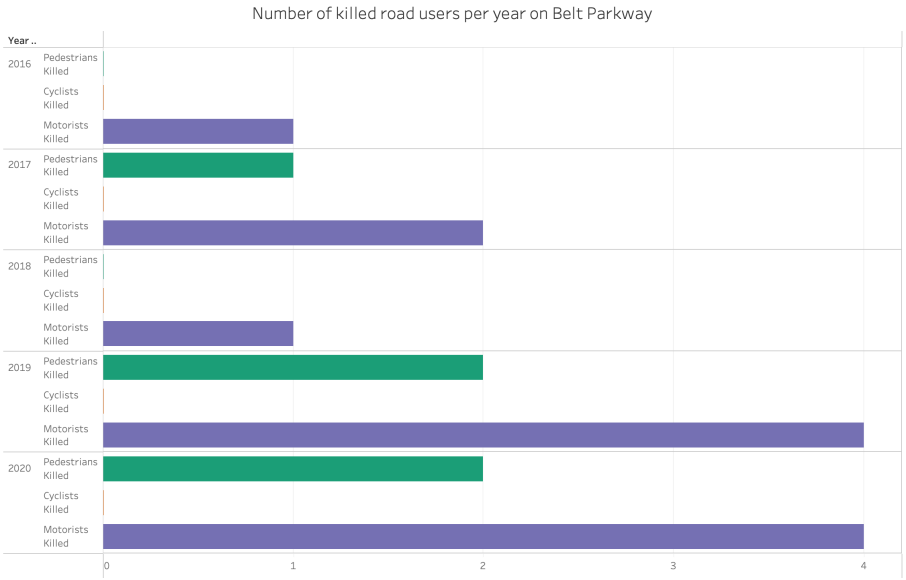
On the other hand, looking at figure 11a, there is a trend in decreasing injured people. Based on this, NYC DOT can predict that the number of injured people will most likely decrease in the next years.

Figure 10b presents the number of killed motorists, pedestrians, and cyclists. It can be seen that the number of killed motorists is meager and has an average of one motorist killed every year. Unlike Belt Parkway, more pedestrians and cyclists are killed in the collisions. In 2016, six pedestrians were killed.

One of the reasons more cyclists and pedestrians are involved in collisions on Broadway than on Belt Parkway is that Broadway is a road where more people travel and walk. Belt Parkway does not, for instance, have pedestrians crossings, and is only allowed for vehicles. This is also a contributing factor to why the number of injured and killed motorists is lower in Broadway - the speed limit is lower, making it less vulnerable to motorists. NYC DOT will, with these analyses, have a clue which is most injured and killed on the two roads and how they should encounter the problem.

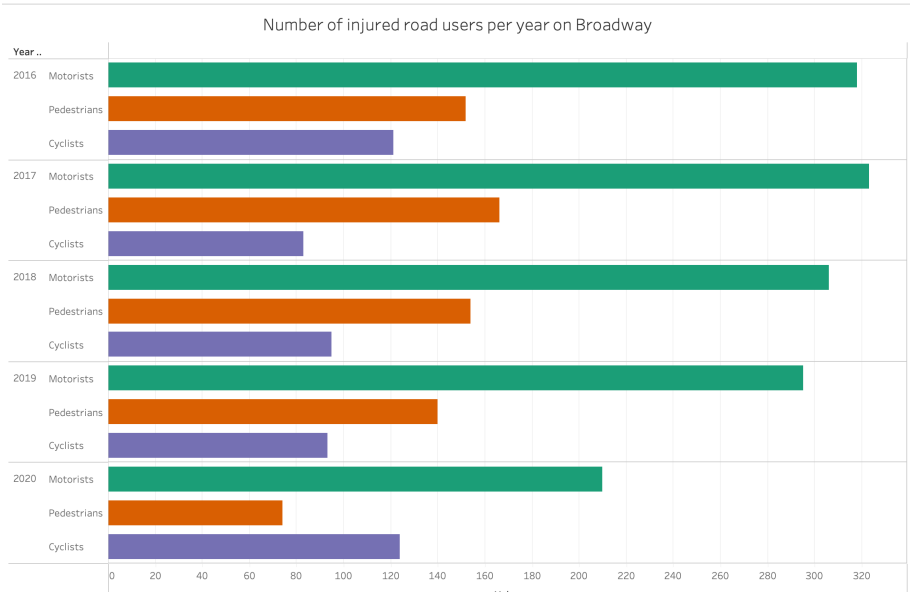


(a) Injured

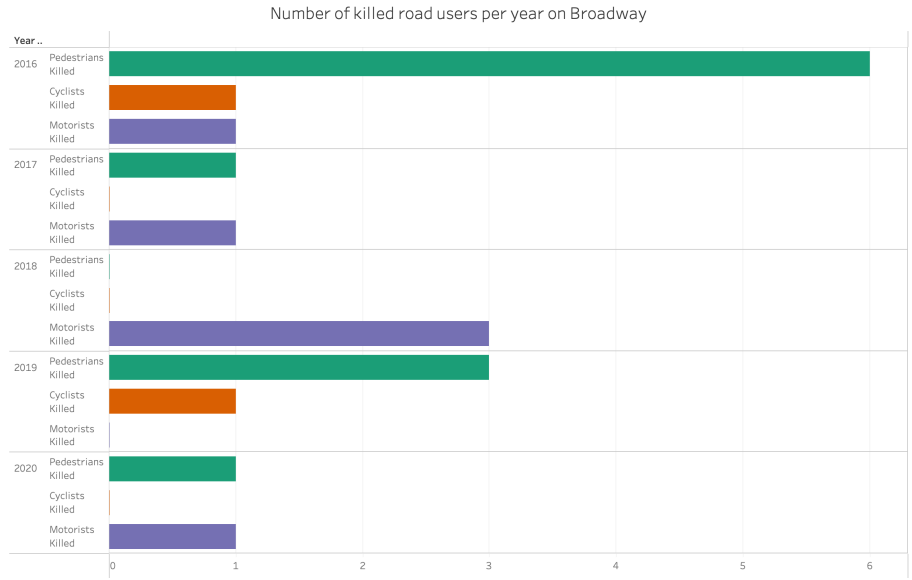


(b) Killed

Figure 10: Number of injured and killed people on Belt Parkway from 2016 to 2020



(a) Injured



(b) Killed

Figure 11: Number of injured and killed people on Broadway from 2016 to 2020

4.2.3 Injury and Death Rate

After analyzing the severity of the collisions in the section above, it is interesting to look at the injury and death rate for Belt Parkway and Broadway compared to all streets. In finding the injury and death rate, the following formula is used for the injury rate:

$$\frac{\text{Number of injured persons}}{\text{Number of collisions}} * 100$$

And for the death rate:

$$\frac{\text{Number of killed persons}}{\text{Number of collisions}} * 100$$

The calculations are filtered based on street names, in this case: Broadway and Belt Parkway. This gives us the number of injured or killed persons per 100 collisions. Table 1 shows the injury and death rate for all streets, Broadway, and Belt Parkway.

Table 1: Injury and death rate for all streets, Broadway and Belt Parkway per 100 collisions

Street	Injury rate	Death rate
All streets	31.8	0.17
Broadway	30.71	0.23
Belt Parkway	41.96	0.13

The death rate and the injury rate provide valuable information to the NYC DOT. The rates give insight into Broadway and Belt Parkway compared to all streets and provide a different view of the two roads. Broadway and Belt Parkway are two roads with lots of collisions, and putting them up towards how many injured or killed people per collision will give a more accurate number of how exposed the road is.

As seen in table 1, the injury rate for all streets is 31.8, and the death rate is 0.17. This means that 31.8 people get injured per 100 collisions, and less than 1 person gets killed every 100 collisions. Since this yields for all streets, these numbers will be used as a comparison throughout the section.

On Broadway, the injury rate is a little lower than for all streets, but the death rate is 0.23, which is slightly higher than in general. In conclusion, the rates for Broadway are the same as for all streets even though it is a highly exposed road. Looking at the numbers for Belt Parkway, the injury rate is 41.96. This is approximately 33% greater than the average. The death rate is at 0.13, which is lower than the average for all streets.

With all this in mind, NYC DOT will have a broader understanding of the situations of these two streets. Maybe NYC DOT should start to focus on Belt Parkway and how to minimize the injury rate as it is way above average, and focus on the death rate on Broadway as this also is above average.

4.3 Decreasing number of collisions based on the most common factors

Reaching the objective of decreasing the number of collisions in New York City first required identifying the most frequent contributing factors. Figure 12 gives an illustration of the most common contributing factors for collisions in NYC from 2016 to 2020.

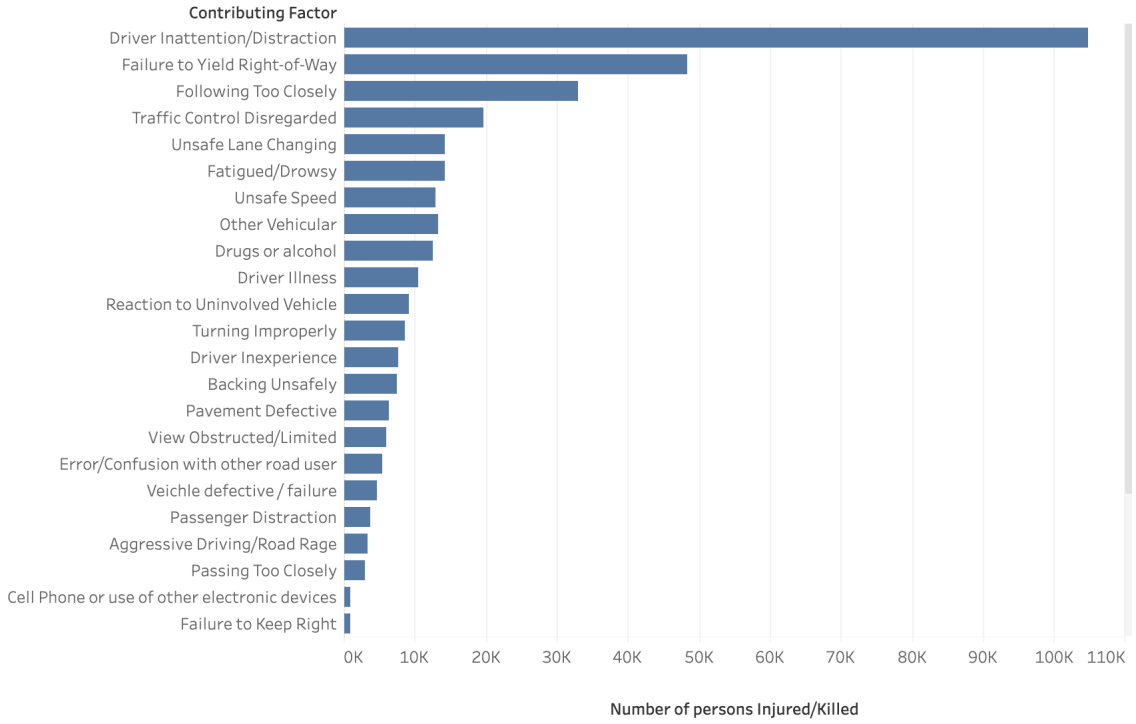


Figure 12: Contributing factors for collisions

To summarize, figure 12 shows that the top six most common contributing factors leading to a collision is: *driver inattention, failure to yield right-of-way, following another car too closely, traffic control disregarded, unsafe lane changing*, or the driver being *fatigued*.

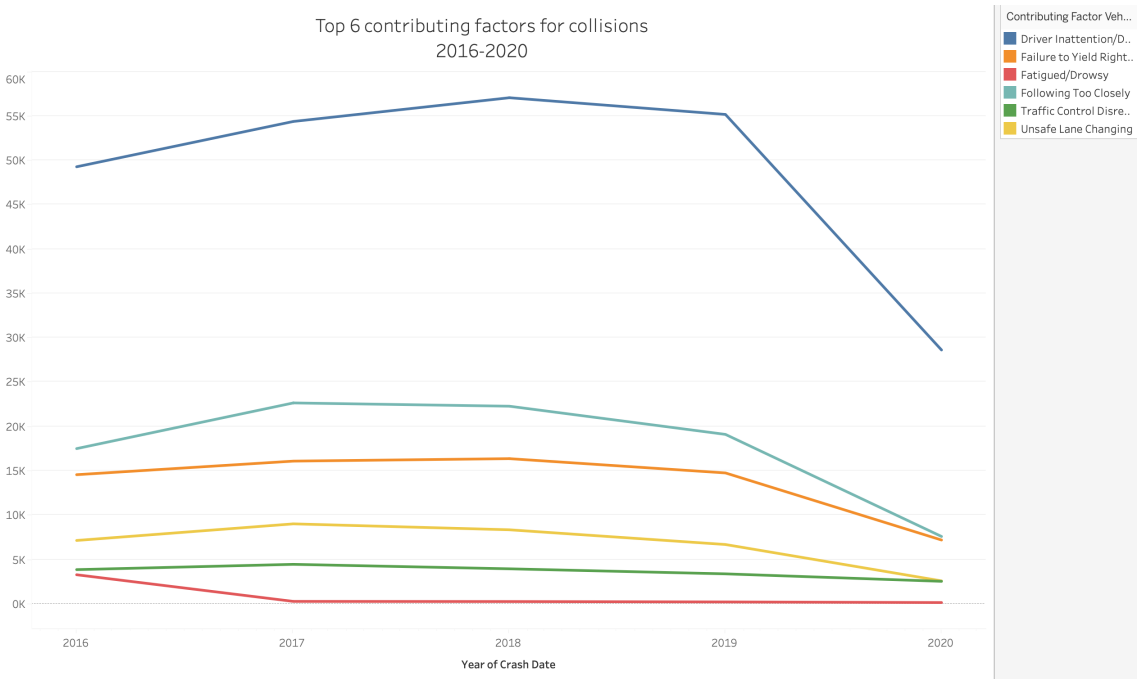


Figure 13: Development of the contributing factors

In figure 13 these factors are shown in a line chart. The chart visualizes that six of the mentioned factors contribute to approximately the same collisions every year except 2020. The last factor, which is due to the driver being fatigued, dropped in 2017. This can be a result of one of Vision Zero's countermeasures for fatigued drivers in 2017 [24]. This measure was made for taxi and FHV (for-hire vehicles) drivers, but the awareness seems to have spread to the regular driver as well. In figure 14 one can see that most of the lines peak in 2018, and abates to an all-time low in 2020. The relative number of collisions caused by the factors included in the figure has decreased to about the same level as in figure 13. This confirms that the awareness has spread to regular drivers.

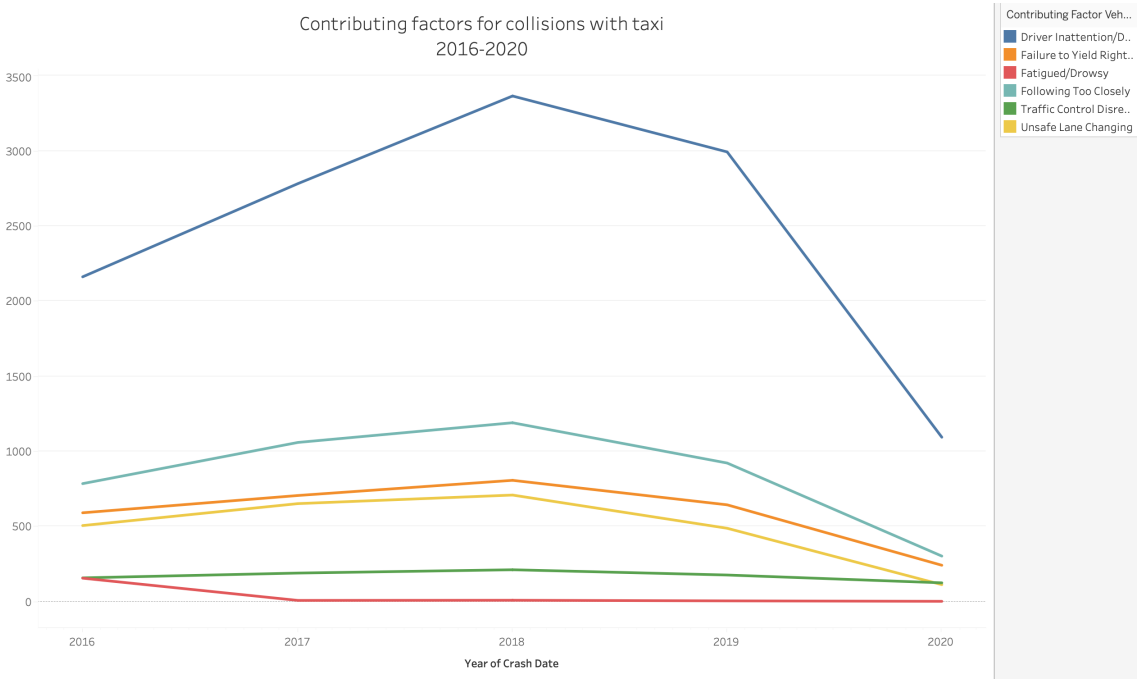


Figure 14: Development of the contributing factors for collisions with a taxi involved

The other mentioned factors will be discussed further in the upcoming sections.

4.3.1 Elaboration of the Contributing Factors

To find appropriate countermeasures, it can be helpful to examine what underlying conditions lead to these factors. Why are so many unable to successfully yield right-of-way? What contributed to unsafe line changing? For instance, failure to yield right-of-way can occur due to a confusing traffic situation or lack of road signs. Unsafe lane changing can occur due to other factors such as the influence of alcohol or drugs or the use of a cellphone or other electronic devices. The most significant contributing factor for collisions is that the driver gets distracted. It could be several reasons for being distracted by the traffic. The most frequent form of distraction while driving has been cell phone use and texting in recent years. Almost 80% of crashes and 65% of near crashes in a study done in the US involved the driver looking away from the roadway just prior to the crash [25]. NYC government Traffic Safety Committee also states that:

- Looking away for two or more seconds will double the risk of a crash or near crash.
- Driver inattention due to drowsiness will increase the risk of a crash or near crash by at least four times.
- A driver who is engaged in a secondary task while driving also increases their risk factor.
- The following actions: talking, listening or dialing a hand-held device; inserting or retrieving a compact disc; operating a PDA; reading, applying makeup or eating will increase the driver risk factor of a crash or near crash by two to three times.

Even though the driver distraction factors lead to a considerable amount of collisions, the factor has decreased in the last years, as seen in figure 13. The severity of this factor is also not that fatal, which will be discussed in section 4.3.2.

Many of the subsequent factors could also have happened in combination with the driver being distracted, making it more important for the NYC DOT to come up with countermeasures for distraction. As stated in section 4.1, the majority of collisions occur during rush hour and 8 AM to 18 PM. The rush hour is, of course, contributing to an extra increase in collisions triggered by distracted drivers. However, this also accounts for the rest of the observed contributing factors except for fatigued driving, which has no significant changes based on the time of the day. This observation can be seen in figure 15.

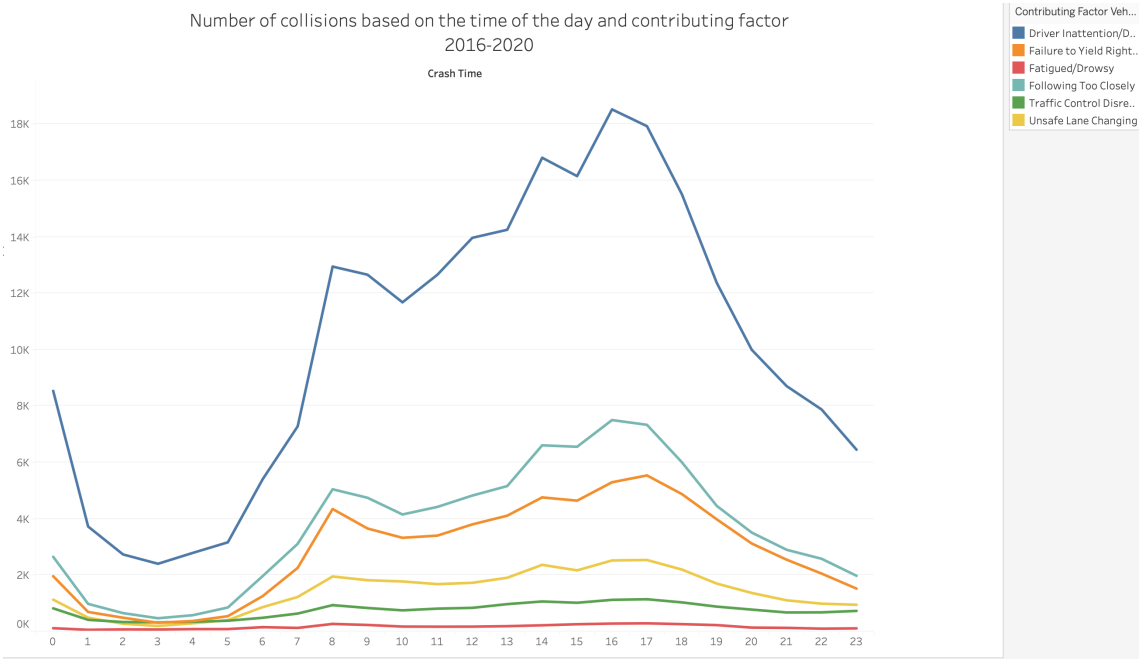


Figure 15: Collisions over time based on the top 6 contributing factors

4.3.2 Severity of the Contributing Factors

Even though distraction is the cause of the accident, the causes of the collision alter depending on the severity of the impact. Figure 16 shows the contributing factors for collisions leading to death. It illustrates how unsafe speed is a major contributing factor in fatal collisions. A fatal collision is one in which one or more people are killed. Two new contributing elements in the top six factors are alcohol involvement and pedestrian/bicyclist error or confusion. It is only logical that unsafe speeds result in fatal crashes because when something goes wrong at high speed, the chances of fatality are great. When a pedestrian or bicyclist gets mixed up in a road with motor vehicles and is confused, the chances of being hit are greater than with lower speeds, and the consequences can therefore be lethal.

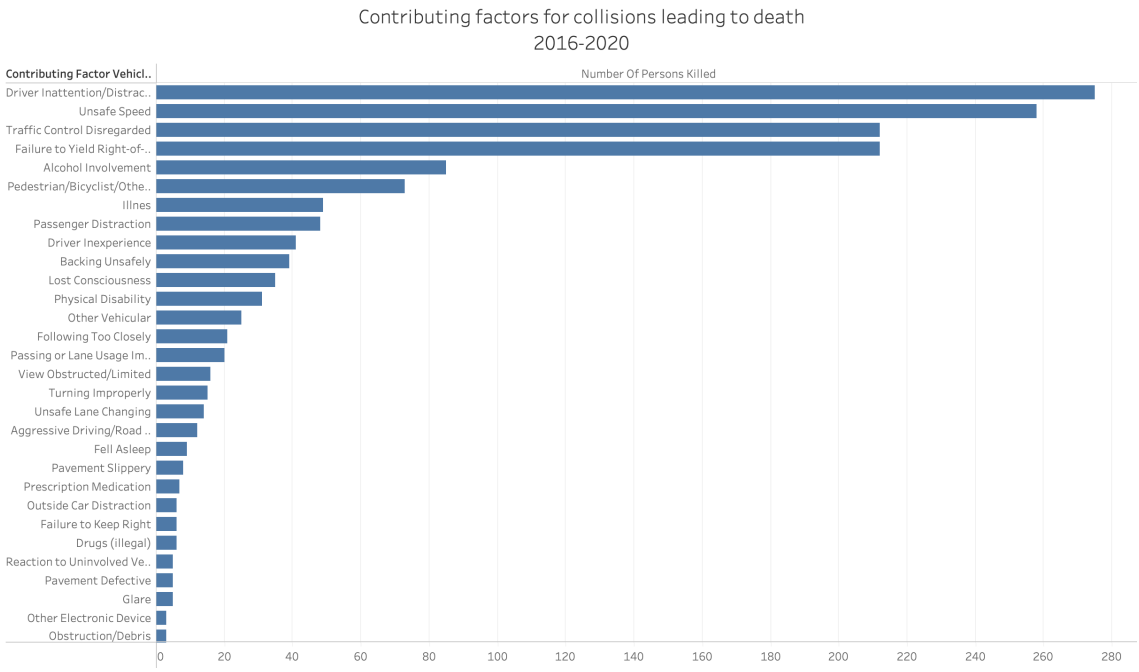


Figure 16: Top contributing factors for collisions leading to death

Watching the trends of these specific factors over the years gives some interesting results. It seems like more people are getting killed due to unsafe speed since 2018. The amount of deaths has increased since 2018. This holds true not only for unsafe speed but the overall amount of killed people in collisions in NYC, which can be seen in figure 18. This is a unique observation because while the number of collisions has decreased, the fatality has not. In recent years, both traffic control disregard and pedestrian/bicycle confusion have increased. However, the number of people killed due to driver distraction and failure to yield the right-of-way has decreased. Looking closely, one might find that the driver distraction factor has only decreased from 2019 to 2020, which could indicate that the pandemic is to blame. If this is the case, the distraction factor will almost certainly increase after the pandemic.

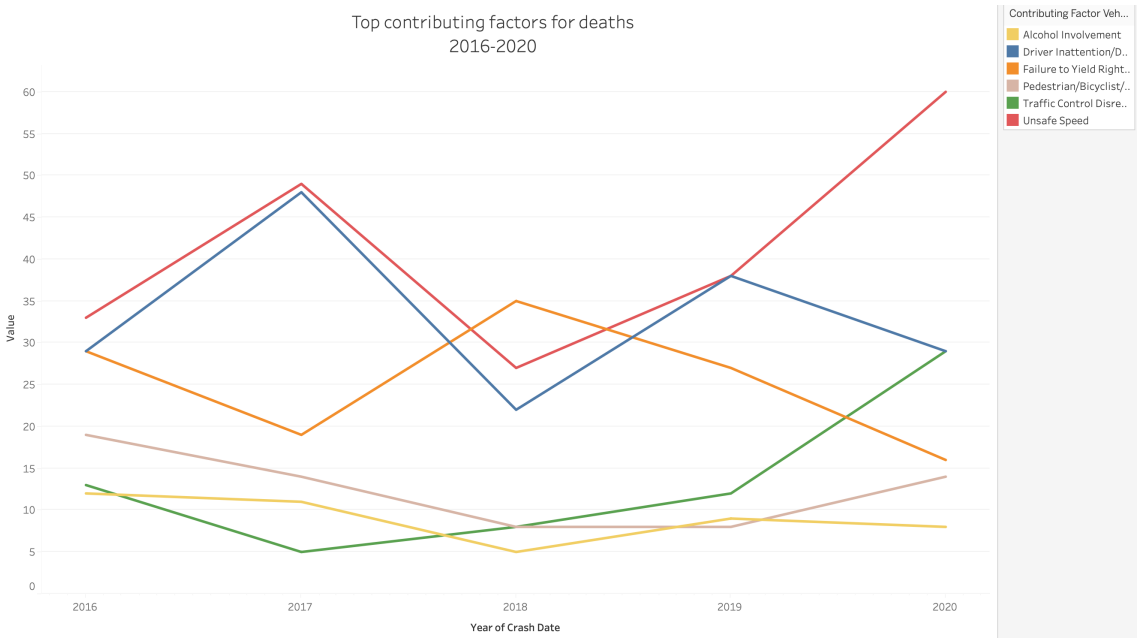


Figure 17: Development of the contributing factors for collisions resulting in deaths

Figure 18 shows the total number of deaths from 2016 through 2020. It is difficult to predict the pattern and comprehend its reasons, primarily because measures were taken after implementing the Vision Zero traffic program.

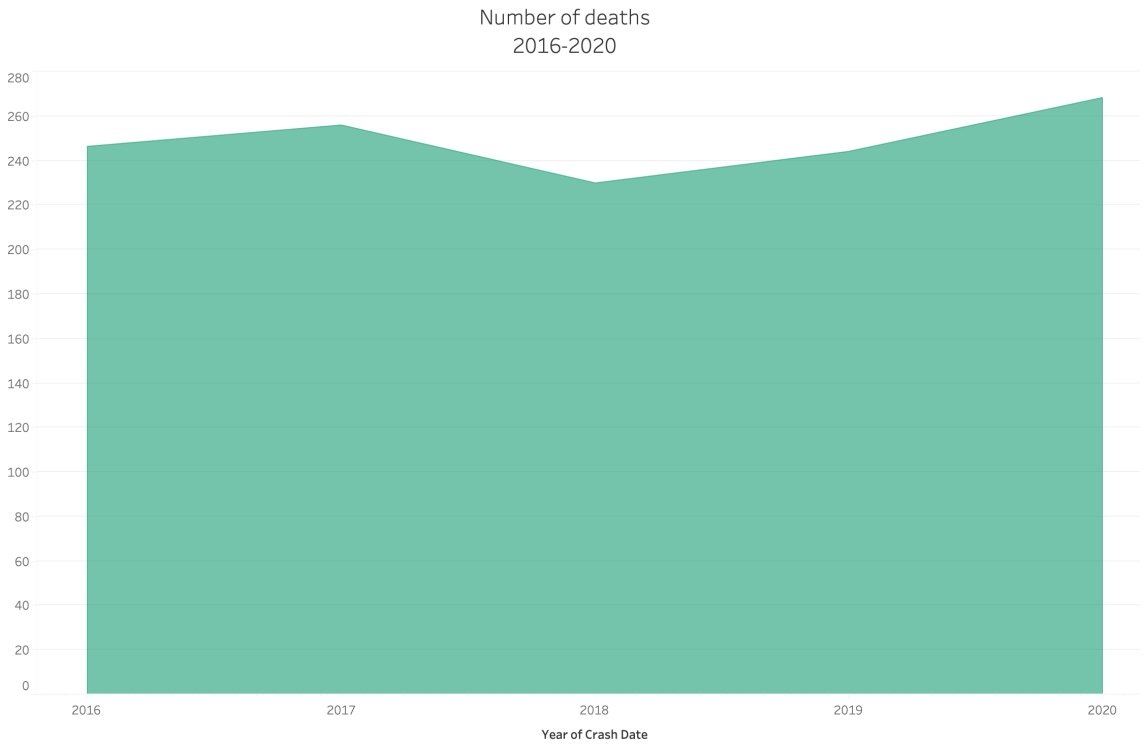


Figure 18: Amount of deaths from 2016 to 2020

5 Limitations

5.1 Data Set Limitations

The Motor Vehicle Collisions data set contains a few limitations. For instance, some crash events contain missing fields, which obviously misrepresents the current situation. For example, some crash events might only include the crash time and street name. This lack of precise data could at worst lead to the results being misleading or deviating from reality.

5.1.1 Unspecified Reason for Collision

A large amount of the data set does not know the reason behind the crash event. More precisely, 36 percent of the data set has "Unspecified" as their main contributor to the crash. Some crashes might be hard to accurately specify the contributing factor as the police come to the aftermath of such a collision. This unknown part of the data set leads to inaccuracy regarding the most common factors behind a collision.

5.1.2 Inconsistencies and Misspellings

In addition, there are some inconsistencies in the data. For instance, some crash events have the same contributing factors but misspellings such as "ilness" and "illness." Furthermore, some streets names are written in different ways and with misspellings. E.g., Belt Parkway and Belt P. In this case, the fields were merged to get more accurate data.

5.1.3 Composite Factors

Certain factors may overlap with more generic ones. For example, 'Fatigued/Drowsy', 'Failure to Yield' or 'Following Too Closely' could be termed 'Driver inattention/Distraction.'. For the police officers that file the reports, it is more convenient to pass something off as a more generic factor, but this decreases the data set's accuracy making it less helpful in taking measures. The more granular the data, the simpler it is to conduct effective analyses.

5.1.4 COVID-19

As one could expect, the number of collisions deviated significantly during the COVID-19 outbreak compared to the rest of the period. This is quite obvious given that the pandemic decreased the overall need for travel because everything was closed. Consequently, the collisions decreased from approximately 17 000 in December 2019 to approximately 4 000 in April 2020. This drastic change might cause miscalculations for their predictions of vehicle collisions in NYC. It is worth noting that this decrease is unlikely to imply that the road network has improved; the total number of trips simply dropped. Overall, the data set is quite helpful in predicting collisions, despite the aforementioned limitations in this data set.

6 Interpretation and Recommendations

6.1 Decreasing Number of Collisions on the Most Exposed Roads

In subsection 4.1, the analysis of the most exposed roads can be summarized into the following:

1. Highways are some of the most exposed roads in New York City. The top four most exposed roads are all highways.
2. Unsafe lane changing and speed are common factors to collisions on highways.
3. The most exposed non-highway road is Broadway in New York City.

6.1.1 Recommendations

The first recommendation to reduce collisions is to station additional officers at the most exposed roads at the most exposed hours. As a collision entry includes both time and location, the NYPD can predict which roads are the most likely for collisions. Installing additional monitors on certain roads serves as a reminder to drivers passing by to drive well. In addition, officers can prevent collisions by punishing reckless drivers and warning distracted ones. The number of road collisions would reduce by removing the most collision-exposed drivers. An officer who can remove the threatening vehicle from the traffic would reduce collisions on the exposed road.

Expanding the most exposed roads to increase the capacity of these roads might be another worthwhile recommendation. Increasing the number of lanes and figuring out how to safely merge lanes might be crucial to reduce collisions.

Lastly, decreasing the number of vehicles on the most exposed roads would be sustainable and effective. In recent times, car ownership in NYC has increased. Reducing the number of vehicles on the road would open up the roads and make it easier for vehicles to navigate. In return, fewer cars also mean fewer collisions. Investing in public transport is essential for reducing the reliance on personal vehicles in NYC.

6.1.2 Implementation Plan

Recommendation	Stakeholders	Timeframe	Success Criteria
Increase the amount of officers patrolling the 4 most exposed roads	Patrolling Officers, Police Department	Jan 2022 - Jan 2023	Decrease the amount of collisions on the 4 most exposed roads by 10%
Expand the 4 most exposed roads	Street and Roadway Construction	Jan 2022 - Jan 2023	Decrease vehicles on the main roads by 15%
Increase the amount of public transport	New York City Transit Authority	Jan 20223 - Jan 2023	Increase the usage of public transport by 15%

6.2 Decreasing Number of Collisions Based on the Most Common Factors/Triggers

The analysis preformed in section 4.3 resulted in three main findings:

1. The most common factor for collisions in NYC is driver inattention, failure to yield right-of-way, following another car too closely, traffic control disregarded, unsafe lane changing, and driver fatigued or drowsiness. However, a lot of the common factors are a composite of several factors.
2. Driver inattention is the most common factor for all types of collisions.
3. The deadliest cause for collisions is unsafe speeding. This factor is increasing.

6.2.1 Recommendations

Road safety and attitude campaigns can be an effective countermeasure for preventing collisions [26]. Such campaigns can affect people to change their attitudes and perception towards traffic safety. According to studies, road safety campaigns have resulted in an overall reduction of road accidents. This is because of reduced speeding among drivers and increased risk comprehension [26].

Road safety campaigns can also be preventative when it comes to reducing accidents caused by driver inattention. In this scenario, the campaigns should focus on educating drivers about what to avoid while driving to retain their attention to the traffic. Multitasking, like eating while driving and mobile usage are all things to avoid while driving. This preventative countermeasure can be applied to several common factors for collisions, for example, teaching drivers about the importance of following speed limits and the consequences of not enforcing the rules.

Increasing the number of speed traps and automated roadside speed cameras might be one approach towards decreasing accidents caused by unsafe speeding. In addition, a possible countermeasure could be to enhance the penalties for violating traffic rules. This can be applied to both speed violation and driver inattention causes.

6.2.2 Implementation Plan

Table 2 lists each of the recommendations, as well as respective stakeholders and a suggested time frame for the implementation of this recommendation. Also, a success criterion for each recommendation is defined.

Table 2: Implementation plan for decreasing the number of collisions based on most common causes

Recommendation	**Stakeholders**	**Time-frame**	**Success criteria**
Increase number of automated roadside speed cameras	Police department	Jan 2022 - Jan 2023	Decrease the number of collisions caused by unsafe speeding by 10%
Create road safety campaigns educating drivers on importance of remaining attention to the traffic	Public relations department	Jan 2022- Jan 2023	Decrease collisions caused by driver inattention by 10%
Increase penalties for violating traffic rules	Police department, The National Highway Traffic Safety Administration	Jan 2022 - Jan 2024	Decrease number of collisions by 10%

6.3 Further Analysis

NYC DOT have already implemented campaigns and countermeasures to prevent collisions, but these have to be followed up, revised and renewed if necessary. To decrease the amount of collisions, NYC DOT needs to do annual analysis to keep coming up with new countermeasures, and see if the current implemented ones have worked. To improve the next iterations, all the limitations have to be considered and fixed. E.g. the data quality can be increased by coursing the police officers in how to fill in data and teach them the importance of good and accurate data. The quality of the data can also be enhanced by digitizing the data collection with forms that has auto fill for e.g. street name and contributing factor. Then the risk of misspelling will decrease.

Including this, the effects of COVID-19 needs to be analyzed to figure out the actual outcome of the pandemic. The collisions might have dropped during the pandemic, but it also might increase afterwards due to more vehicles on the road.

The next step for NYC DOT now is to implement the recommendation from this report, collect more accurate data, and finally do annual iteration to find countermeasures. This will hopefully improve the current traffic situation of NYC.

References

[1] Marian White. *The Top 10 Largest U.S. Cities by Population*. 2021. URL: https://www.moving.com/tips/the-top-10-largest-us-cities-by-population/ (visited on 10/04/2021).

[2] INRIX. *2021 Urban Mobility Report*. 2021. URL: https://static.tti.tamu.edu/tti.tamu.edu/documents/mobility-report-2021.pdf (visited on 10/04/2021).

[3] World Health Organization. *Road Traffic Injuries*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (visited on 10/04/2021).

[4] New York Police Department (NYPD). *Motor Vehicle Collisions - Crashes*. Updated October 2, 2021. URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

[5] Fawcett T. Provost F. *Data Science for Business*. OReilly, 2013.

[6] OECD. *The Path to Becoming a Data-Driven Public Sector*. OECD, 2019.

[7] *Award: Compstat: A Crime Reduction Management Tool*. 2021. URL: https://ash.harvard.edu/news/compstat-crime-reduction-management-tool (visited on 10/06/2021).

[8] *Motor Vehicle Collisions - Crashes*. 2021. URL: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 (visited on 10/13/2021).

[9] *Analysis of New York City Motor Vehicles Collisions*. 2021. URL: https://towardsdatascience.com/analysis-of-new-york-city-motor-vehicles-collisions-927da110dfc7 (visited on 10/13/2021).

[10] *Fakta om bilbelte*. 2021. URL: https://www.vegvesen.no/trafikkinformasjon/trafikksikkerhet/kampanjer/bilbelte/fakta-om-bilbelte/ (visited on 10/13/2021).

[11] https://analyticsindiamag.com/. *CRISP-DM pros and cons*. 2020. URL: https://analyticsindiamag.com/why-is-crisp-dm-gaining-grounds/ (visited on 10/04/2021).

[12] Giles A. Hindle and Richard Vidgen. *CRISP-DM pros and cons*. 2017. URL: https://isiarticles.com/bundles/Article/pre/pdf/161699.pdf (visited on 10/04/2021).

[13] www.datascience-pm.com. *CRISP-DM pros and cons*. 2020. URL: https://www.datascience-pm.com/crisp-dm-2/ (visited on 10/04/2021).

[14] Patrick Mikalef. *Thinking like a data scientist*. University Lecture. 2021.

[15] Mithun Sridharan. *CRISP-DM: A framework for Data Mining & Analysis*. 2018. URL: https://thinkinsights.net/digital/crisp-dm/ (visited on 10/18/2021).

[16] Mar. 2015. URL: https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond.

[17] Wikipedia. *Survivorship bias*. 2021. URL: https://en.wikipedia.org/wiki/Survivorship_bias (visited on 11/08/2021).

[18] Patrick Mikalef. *Data Visualization - Principles of design & Power BI*. University Lecture. 2021.

[19] Patrick Mikalef. *Data visualization: Sell your story*. University Lecture. 2021.

[20] Mark Harrower Cynthia Brewer and The Pennsylvania State University. *Color Brewer 2*. URL: https://colorbrewer2.org/ (visited on 11/01/2021).

[21] https://www.eastcoastroads.com. *Belt Parkway*. 2020. URL: https://www.eastcoastroads.com/states/ny/parkways/belt (visited on 10/18/2021).

[22] https://www.broadwayleague.com. *Broadway Season Statistics*. 2020. URL: https://www.broadwayleague.com/research/statistics-broadway-nyc/ (visited on 10/18/2021).

[23] CBS New York Team. *15 People Hospitalized after Pileup on Belt Parkway*. 2021. URL: https://newyork.cbslocal.com/2021/09/20/belt-parkway-crash/ (visited on 10/20/2021).

[24] NYC Government. *Fatigued Driver prevention*. 2021. URL: https://www1.nyc.gov/site/tlc/about/fatigued-driving-prevention.page (visited on 11/03/2021).

[25] Our World in Data. *Distracted driving*. 2021. URL: https://trafficsafety.ny.gov/distracted-driving-1 (visited on 11/03/2021).

[26] Traffic Injury Research Foundation. *ROAD SAFETY CAMPAIGNS - WHAT THE RESEARCH TELLS US*. 2015. URL: https://tirf.ca/wp-content/uploads/2017/01/2015_RoadSafetyCampaigns_Report_2.pdf (visited on 11/02/2021).