2026

COVID Drivers

DID AGGRESSIVE DRIVING BEHAVIORS INCREASE AFTER COVID?

LISA OVER

Table of Contents

Section 1: Business Understanding

1.1 Background

Organizational/Context Description

PA Data Discourse (PDD) is a new digital news publication focused on investigative public interest reporting. PDD collects and analyzes open data from state, county, and municipal portals to provide data-driven insights on topics of interest to Pennsylvania residents.

Domain and Context Assessment

A current topic of interest that has open questions involves PA drivers and driving behaviors that may have changed during and after the COVID-19 pandemic. A majority of Americans believe drivers have been more aggressive since the lockdown.

A Pew Research Center (Leppert, 2024, November 12) report published in November 2024 reveals that a majority of Americans believe more drivers have been distracted, aggressive, and/or under the influence of drugs or alcohol since the pandemic. As many as 78% of Americans say more people are distracted by their cell phones, 63% say more people are driving too fast, 63% say more people are driving aggressively (weaving, tailgating, or running red lights), and 51% say more people are driving while under the influence of alcohol.

A National Safety Council (NSC, 2020, July 21) report published in the summer of 2020 shows that although there were far fewer cars on the road during the pandemic, the fatality rate per miles driven rose by 23.5% during the quarantine in May 2020 compared to the same month the previous year. PennLive reporter Jordan Wolman interviewed the manager of statistics for the NSC, Ken Kolosh, who explained that there was "one death for every 84 million miles driven in May 2019, but one death for every 68 million miles driven in May 2020." (Wolman, 2020, July 30) The *Pennsylvania Capitol-Star* (Henderson, 2023, November 13) reported a 12% rise in traffic fatalities in Pennsylvania from 2019 to 2022 despite fewer cars on the roads in the state.

The Foundation for Traffic Safety (FTS) of the American Automobile Association (AAA) (Tefft et al., 2022) conducted a Traffic Safety Culture Index survey in 2020 where respondents reported the amount of time they have spent driving during the first few months of the pandemic and if they engage in risky driving behaviors, such as texting, speeding, drinking, and running red lights. The results of this study show that drivers who engage in risky behaviors increased the amount of time they spent driving while safer drivers decreased the amount of time they spent driving. Another FTS study (Tefft et al., 2021) involved a New American Driving Survey where respondents reported their travel from the previous day. The goal of the study was to quantify the reduction in the amount of driving during the pandemic, and the results showed that the number of trips taken by U.S. residents dropped abruptly in April 2020. Compared to the average number of daily trips from July through December in 2019, the average number of daily trips dropped by 40% in April 2020.

Problem Situation

The studies outlined above confirm the following statements about driving habits in the United States during and after the pandemic:

- Americans believe aggressive driving has increased since the COVID-19 pandemic, and super majorities believe that one or more of the following specific behaviors have increased: distracted driving, driving too fast, and reckless driving including weaving, tailgating, and running red lights. (Leppert, 2024, November 12)
- Fatalities per miles driven in May 2020 increased significantly over the same month in 2019. (NSC, 2020, July 21)
- Americans who admit to engaging in risky driving behaviors increased the amount of time they spent driving while safer drivers reduced the amount of time they spent driving during the first few months of the pandemic. (Tefft et al., 2022)
- The average number of daily trips taken by car dropped significantly in April 2020 during the COVID-19 lockdown compared to the last six months of 2019. (Tefft et al., 2021)

Data Science Justification

While these studies confirm important trends and patterns, none of them show if aggressive driving actually increased during COVID or if it has since declined or stabilized. They also do not show if the rate of accidents caused by at least one aggressive driver has changed or if any type of aggression (speeding, distraction, recklessness) has increased more than others. Data scientists can answer these questions using appropriate data and modeling techniques.

It is important for the public to know if they are correct about an increase in aggressive driving, and it is important to understand which types may have increased. Public awareness brings solutions by moving the public to elect public officials who will enact laws or by moving organizations and individuals to organize and fund advocacy or education efforts.

Stakeholder Identification

The primary stakeholders in this research are the editor and the public. The editor determines if the content is accurate and meets editorial standards, if the methodology is sound, and if the article is compelling and fits within the mission of the publication. The residents of Pennsylvania are concerned about aggressive driving and want to know if their perceptions are correct. This project will answer this question to promote meaningful dialog about driving behavior.

The secondary stakeholders are public officials, groups, and organizations who will be moved by the public to facilitate change through policy, laws, advocacy, and education.

1.2 Project Objectives and Success Criteria

Primary Objective

Determine whether there has been a statistically significant increase in aggressive driving since the COVID-19 pandemic to provide evidence-based conclusions for PA Data Discourse readers about driving risks within 12 weeks.

Secondary Objectives

- Identify specific aggressive driving behaviors that increased during COVID-19.

- Identify co-factors for crashes involving aggressive driving, such as age, day of week, geography, terrain, or atmospheric conditions.
- Develop accessible visualizations for public understanding.
- Create reproducible methodology for ongoing aggressive driving related crash monitoring.

Quantitative Success Criteria

To ensure technical rigor, this project will assess the performance of two change point detection (CBD) algorithms. The performance of each algorithm will be assessed using performance measures, statistical significance thresholds, and confidence intervals that are appropriate for detecting change points given the characteristics of the data and the algorithm.

Qualitative Success Criteria

- Editorial confidence in methodology
- Reader engagement demonstrates understanding
- Public servants, including elected officials and law enforcement, reference the findings
- Methodology adopted by other researchers

The findings will be reviewed by peers and editorial staff before publication. Community engagement will be encouraged and monitored for feedback.

Impact Assessment

- Evidence informing driving decisions for 13+ million residents of Pennsylvania
- Foundation for policy, advocacy, and education effecting the entire commonwealth
- Establishes PA Data Discourse's data journalism credibility
- Creates lasting aggressive driving monitoring capability

Timeline and Milestones

Weeks 1-4 Data integration and assessment
Weeks 5-8 Statistical analysis and model results
Week 9 Draft of paper and report complete
Week 10 Peer review and refinement
Weeks 11 Final production
Week 12 Final publication
Week 13+ Impact monitoring

1.3 Assessment of Responsibility

Legal and Regulatory Considerations

The Pennsylvania Department of Transportation (PennDOT) maintains comprehensive crash data covering all reportable traffic incidents throughout the commonwealth. Data for this project were obtained from the public download section (*PennDOT Crash Data Download*, n.d.) of the Pennsylvania Crash Information Tool (*Pennsylvania Crash Information Tool*, n.d.). No regulations, terms of use, or licensing information is specified for using data from this source. Nevertheless, the principles from the CLeAR Documentation Framework for AI Transparency

(Chmielinski et al., 2024, May) and Data Feminism (D'Ignazio and Klein, 2020) will be adopted to ensure responsible data science practices are implemented throughout this project.
This project adheres to the CLeAR Documentation Framework for AI Transparency as follows:

- *Comparable:* The CRISP-DM Framework is a data mining methodology that provides a structured approach for transforming business problems into data driven solutions and specifies clear documentation guidelines throughout the project life cycle. This project follows CRISP-DM and can be compared to any other project that follows CRISP-DM.
- Legible: Documentation for this project includes the components specified in the CRISP-DM Framework that will be reviewed by editorial staff and an article for a general public audience. Both audiences consist of non-technical readers. The documentation and article will be written to these audiences.
- *Actionable:* The article will provide information and a platform for discussing the results.
- *Robust:* The plan for this project is to continue incorporating data as it becomes available so stakeholders can see if the trends change over time or in response to new policies, laws, advocacy, and education.

Principles from Data Feminism that will influence this project include justice, oppression, equity, co-liberation, reflexivity, and understanding history, culture, and context. (D'Ignazio and Klein, 2020, p. 60)

Privacy and Data Protection

Although the crash data are anonymized, an analysis of the geographic location of crashes could stigmatize certain communities or neighborhoods, and this will affect the primary stakeholders if they live in an area where the aggressive driving rate is high.

Bias and Fairness Concerns

It is possible that the available data are biased because of implicit biases held by law enforcement and/or witnesses. Events are interpreted and information is recorded by fallible human beings.

Stakeholder Impact Analysis

It is imperative to be mindful of the fact that each record in the crash dataset represents one or more people who witnessed or experienced an event that was an inconvenience at its best and traumatic or fatal at its worst. This work will promote traffic safety and is not a venue for sensationalizing dangerous driving behaviors.

Mitigation Strategies

The data scientist recognizes her position of privilege and power, and acknowledges that data are not objective and that avoiding bias requires reflection, listening, understanding, and thoughtful engagement. The focus of this project is on behavioral patterns around traffic safety. The goal is to answer specific questions related to a change in overall behavior. It may be useful to aggregate by county or urban/rural areas but not to specific municipalities or neighborhoods or by socioeconomic factors. Findings will be presented with clarity and appropriate context to avoid misinterpretation.

1.4 Data Science Goals and Success Criteria

Technical Problem Framing

Crash data are collected and timestamped based on the date of the crash. The level of granularity of the public dataset is by month and year. This allows for visualizing the data over time to show any abrupt changes in the data and for applying machine learning techniques for detecting if an abrupt change is statistically significant. The techniques chosen for this project include both supervised and unsupervised learning algorithms. For supervised learning, each record will be coded manually as being before or after COVID-19. For unsupervised learning, the algorithm will find any abrupt change points without this manual coding.

According to Aminikhanghahi and Cook (2017), change point analysis is appropriate for analyzing human activity. Although the example they provide is related to sensor data from mobile devices, the goal of this project is to identify changes in human behavior using data that were collected manually by law enforcement. This project aims to identify if a change point exists between crashes that occurred before COVID-19 and those that occurred after where the property of aggressive driving changed, specifically, that aggressive driving increased. NHTSA (n.d.) defines aggressive driving as follows:

> ...violations that encroach on others' safe space, such as driving much faster than prevailing speeds, following too closely, making unsafe lane changes, and running red lights, either on one occasion or over a period of time, may indicate a pattern of aggressive driving. (para. 1)

This project uses variables related to these behaviors in addition to variables related to anything that existed prior to the crash, including other vehicles on the road that ended up in the crash, child passengers who may have distracted the driver, the age of drivers involved, drugged driving, drunk driving, road and weather conditions, among others.

Data Science Objectives

Establish statistically rigorous change point detection on observations of crash data that is suitable for public communication. Analysis must isolate crashes from geographic, atmospheric, demographic, and socioeconomic confounders and provide interpretable visualizations and narrative. For example, 'the likelihood that the aggressive driving rate increased after COVID-19 is X' and 'the odds of a crash involving at least one aggressive driver increase to X after COVID-19.'

Establish statistically rigorous evaluation of aggressive driving behaviors on observations of crash data that is suitable for public communication. Analysis must identify aggressive driving behaviors that increased during COVID and determine which, if any, continued to be high or to increase since COVID. For example, 'Driving while X or using X increased during COVID and continues to increase or has stabilized at a high level.'

Technical Success Criteria

Detect if a change point exists during COVID-19, determine if the nature and direction of the change indicates an increase in aggressive driving behaviors, and determine the percentage increase in NHTSA aggressive driving behaviors. Change point detection will be confirmed by a confusion matrix and with measures of performance for classification, including accuracy, sensitivity, specificity, G-mean, F-measure, and Area Under the Curve (AUC). The model must perform with G-mean, F-measure, and AUC each greater than 0.7.

Analytical Approach Overview

Compare the performance of the following algorithms in detecting abrupt changes in aggressive driving rate and in determining the nature and magnitude of that change for specific behaviors:

- ARIMA Model – fit an ARIMA model to forecast future values and to compare observed and predicted values for divergence. (WANGPRATHAM, January 31, 2023)
- Random Forest Model – fit a random forest model to identify significant deviations between the predicted and observed values, which, if found, indicate significant changes in the trend or pattern that would be a change point. (WANGPRATHAM, January 31, 2023)

Objective Technical Mapping

Statistically significant G-mean, F-measure, and AUC

Credible evidence to support public opinion that people have been driving more aggressively since COVID-19

Article reaches 60% of Pennsylvania residents

Informed public participates in advocacy and education efforts

Reduction in crashes where aggressive driving played a role.

Constraints and Assumptions

The crash data are assumed to be reported accurately. It may not be possible to control for all geographic, atmospheric, demographic, and socioeconomic factors.

Naive Bayes and Logistic Regression do not require the time series to be independent and identically distributed. However, the number of within-state sequences will greatly outnumber the change point sequences, causing imbalanced data. (Aminikhanghahi and Cook, 2017)

Some of the models may not be easy to explain to a non-technical audience. A Random Forest Model may provide the least technical explanation.

1.5 Project Plan

Phase-specific Planning

- Business Understanding
  - Evaluate the project description and *PennDOT Open Data Portal Crash Data Dictionary and Field Constraints Tables* document.
  - Review the CLeAR Documentation Framework and principles from Data Feminism.
  - Develop an understanding of stakeholders.
- Data Understanding
  - Download the data and determine which variables to keep for analysis.
  - Create new variables: timestamp, binary flag for pre/post COVID, and text variables from encoded categorical variables to make it easier to review.

    - Visualize univariate and bivariate relationships, specifically, create line plots to view the trends of each binary flag variable selected for the project.
- Data Preparation
    - Prepare the data for ARIMA and Random Forest modeling.
    - Create a dataset with the timestamp as an index.
    - Decompose the data into trend, seasons, cycles, and noise.
- Modeling
    - Run several models with single variables, including aggressive driving flag and specific aggressive driving behaviors.
    - Run models to analyze differences among counties or between urban and rural areas.
    - Run models to analyze differences between driver ages.
- Evaluation
    - Review and compare the models.
    - Write the final report and paper.

Timeline and Milestones

Task	Date Range	Deliverable
Business Understanding	January 19 – February 2	Business and Data Understanding Report
Data Understanding	February 3 – February 16	Business and Data Understanding Report
Data Preparation	February 17 – March 2	Data Preparation and Modeling Report
Modeling	March 3 – March 23	Data Preparation and Modeling Report
Evaluation	March 24 – May 4	Final Report

Risk Identification and Contingencies

There is a risk that the data will be poor quality with inconsistencies and logic errors among related fields and/or missing data.

The *PennDOT Open Data Portal Crash Data Dictionary and Field Constraints Tables* document states that the public crash data consists of data obtained from multiple sources, including investigating police agencies who submitted a Police Crash Report, drivers who submitted the AA600 Driver's Accident Report form, and calculated fields from the Crash Data Analysis and Retrieval Tool (CDART) database. The datasets compiled from these sources may have some errors and inconsistencies as noted in other sections. (PENNDOT, 2025)

Variables that have to many missing values or logic errors will be removed from the dataset before analysis.

Section 2: Data Understanding

2.1 Data Inventory

Data for this project are from public datasets about crashes in the Commonwealth of Pennsylvania from 2005 through 2024. For each year, eight comma separated values (CSV) files with tabular data reside in folders named Statewide_<year>. The eight CSV files are named as follows:

- COMMVEH_<year>.csv

- CRASH_<year>.csv
- CYCLE_<year>.csv
- FLAGS_<year>.csv
- PERSON_<year>.csv
- ROADWAY_<year>.csv
- TRAILVEH_<year>.csv
- VEHICLE_<year>.csv

Data Collection Methodology

The MDS course instructors used a Python script to download the data from the *PennDOT Pennsylvania Crash Information Tool* (*Pennsylvania Crash Information Tool*, n.d.) and arranged the Statewide_<year> folders inside a 'data/raw' folder hierarchy. The course instructors shared their download program, *download-data.py*, which is located in the top-level 'data' directory. The zipped data folder was downloaded from the course website for use in the project.

The *PennDOT Pennsylvania Crash Information Tool* allows users to download data manually with no permissions or authentication required.

No problems were encountered in downloading, unzipping, and accessing the data.

2.2 Data Description

Comprehensive Data Dictionary

Appendix A: Data Dictionary for CRASH and *Appendix B: Data Dictionary for FLAGS* show the business description, data type, count of unique values, range (if numeric) or list of unique values (if text), and count and percentage of missing values for each variable selected for this project.

Appendix E: File Information displays the year, file path, file size, number of rows, and number of columns for each dataset.

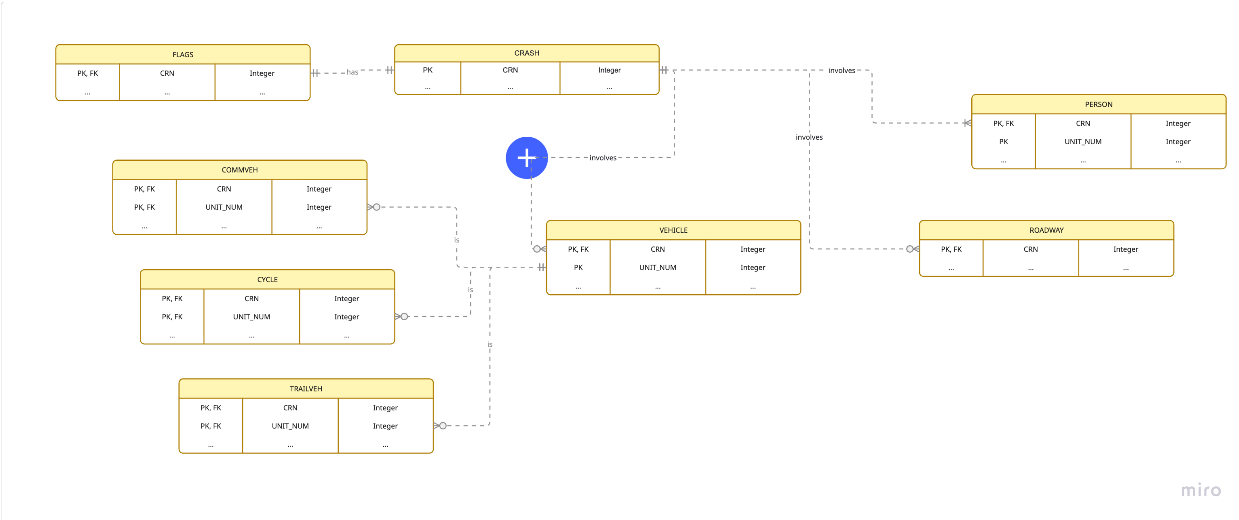
Appendix F: Columns and Data Types displays the dataset category, e.g., COMMVEH, CRASH, CYCLE, etc. and data type for each column. These tables include all of the data to show the total scope, however, only a select number of variables relevant to the business objectives will be used for the project.

Data Structure and Relationships

Figure 1: High-level Structure of Data Diagram shows the relationship between the CRASH and FLAGS datasets. CRASH contains the unique identifier and primary key Crash Record Number (CRN), and FLAGS contains the unique identifier and foreign key CRN. There is a one-to-one relationship between CRASH and FLAGS on CRN. ***Error! Reference source not found.*** and ***Error! Reference source not found.*** provide the business meaning (PENNDOT, 2025), data type, range (for numeric variables) or unique values (for text variables), primary and foreign keys, and unique identifiers for each selected field in the CRASH and FLAGS datasets, respectively.

Figure 1: High-level Structure of Data Diagram shows the relationships between the 8 sets of files.

Figure 1: High-level Structure of Data Diagram



Temporal Coverage and Granularity

Data for this project are from public datasets about crashes in the Commonwealth of Pennsylvania from 2005 through 2024. The year and month of a crash is the temporal frequency of the public dataset. The specific day on which the crash occurred is missing, and the time the police arrived is missing from most records.

Data Volume and Scale

All 160 datasets (8 files for each year over 20 years) require 4.18 GB of disk space for storage. The CRASH dataset has 2,461,193 records.

For this project, selected data include variables from the CRASH and FLAGS datasets that are associated with factors that existed before the crash. For example, the aggressive driving variables, vehicle type indicators, school and work zone indicators, driver age counts, weather, location type, and road conditions are included. Variables that are related to injuries, fatalities, number of vehicles involved, changed traffic patterns after the crash, and crash site cleanup are excluded. These variables would answer questions related to the severity of crashes, but those questions do not relate to the business objectives. The CRASH and FLAGS datasets have enough information to control for age (YOUNG_DRIVER and DRIVER_COUNT_<age>), road conditions (ROAD_CONDITION), location type (LOCATION_TYPE), and county or urban vs. rural (COUNTY and URBAN_RURAL).

Categorical Variable Encoding

The following CRASH fields are coded or have abbreviations. See *Appendix C: Coded CRASH Variable Fields* and *Appendix D: Enumeration of Counties* for the definitions and full values for each field.

- COUNTY
- DAY_OF_WEEK
- ILLUMINATION
- INTERSECT_TYPE

- LOCATION_TYPE
- ROAD_CONDITION
- URBAN_RURAL
- WEATHER1
- WEATHER2

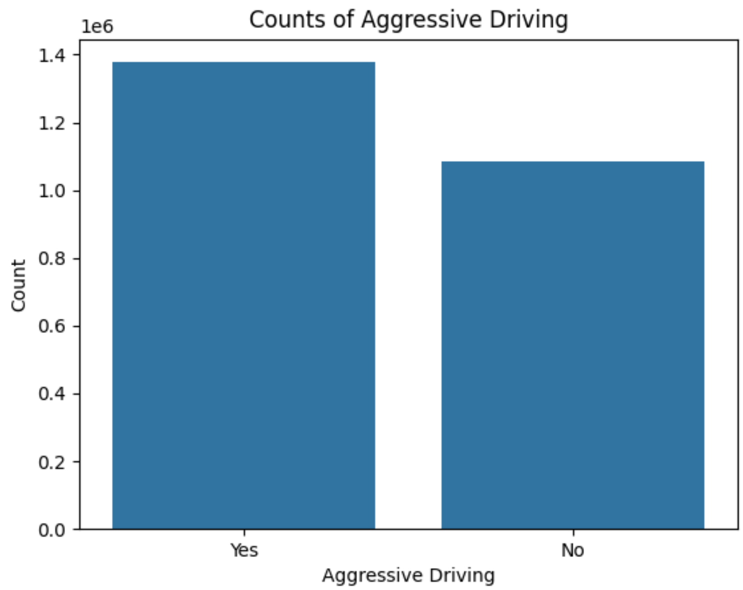
2.3 Data Exploration

New variables related to the crash date, month, and day were created for use in the exploration.

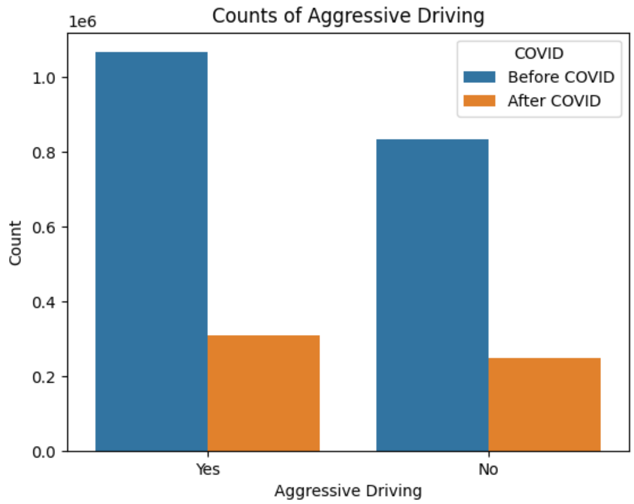
- CRASH_DATE is a timestamp variable created using the crash month and year. The day of the month was set to 01. This variable is used for grouping and displaying time series plots.
- CRASH_MN_NAME is a text variable derived from the numeric month to have a readable month to display on any visualizations.
- POST_COVID is a binary (0, 1) variable where 0=before COVID and 1=after COVID. Dates before January 1, 2020 were marked as before COVID and dates on or after January 1, 2020 were marked as after COVID. The chosen cut-off for pre- and post-COVID may need to be adjusted.

Univariate Analysis

The following count plot of the distribution of AGGRESSIVE_DRIVING, shows that more crashes occur when one or more aggressive driving behaviors were present than when no aggressive driving behaviors were.



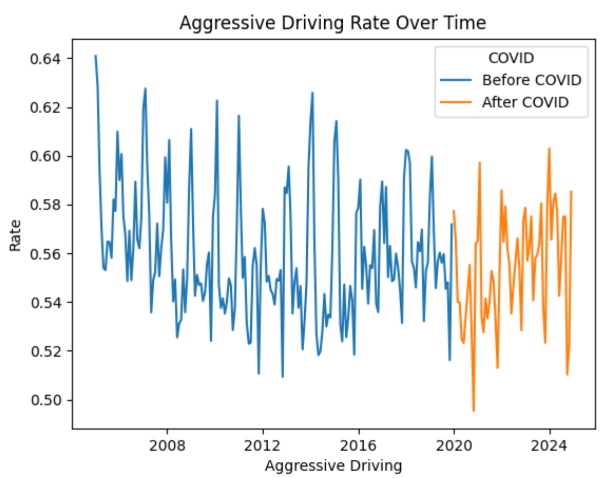
The following count plot of the distribution of AGGRESSIVE_DRIVING whetheer the crash occurred before or after COVID shows how unbalanced the dataset is with many more crash records from before COVID (2005 through 2019) than those after (2020 through 2024). The trend from 2005 - 2024 is important to study so these counts were turned into rates to visualize the time series in the next section.



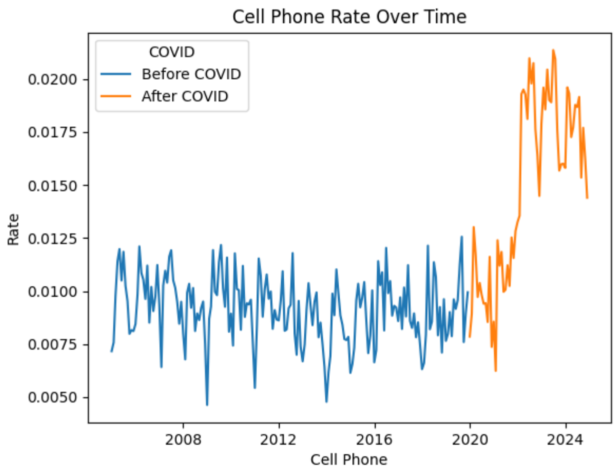
Bivariate Relationships with Temporal Patterns and Trends

The rate of aggressive driving overall and of specific behaviors was calculated as the number of crashes where aggressive driving was exhibited divided by the total number of crashes. This was aggregated monthly.

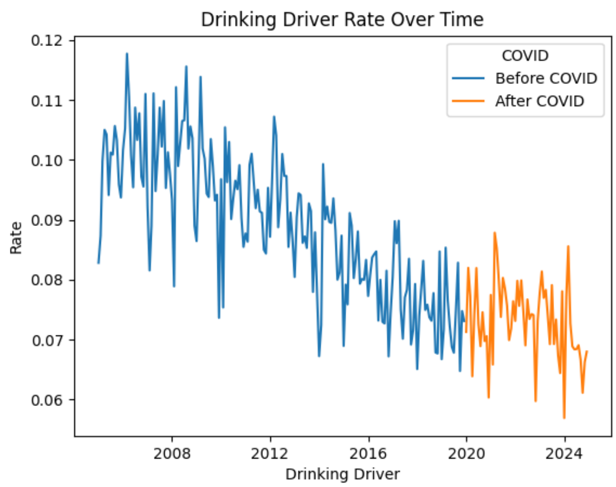
There is an AGGRESSIVE_DRIVING variable in the flags dataset that indicates that one or more aggressive driving behaviors were present during the crash. There are also variables for several specific types of aggressive driving behaviors. The following figures show some of these behaviors over time. The components of the time series including trend, seasonality, cyclicality, and noise, will be evaluated later.



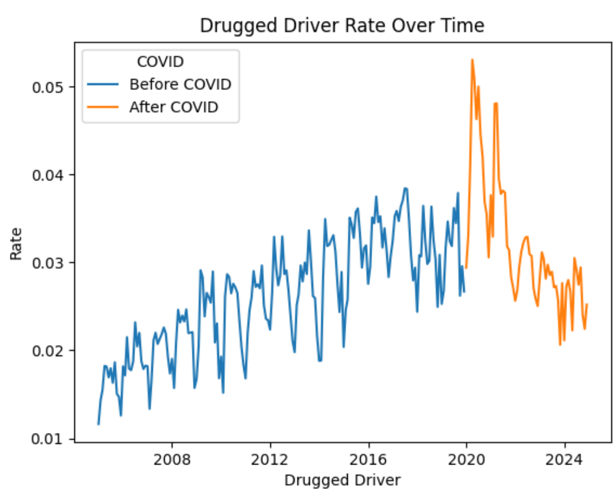
When considering aggressive driving as a single action that includes one or more specific aggressive behaviors, it appears that aggressive driving does not increase with COVID.



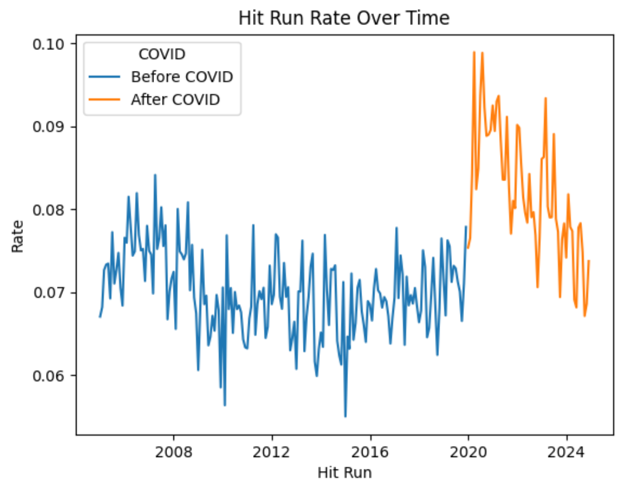
The use of cell phones while driving spiked after COVID between 2022 and 2023.



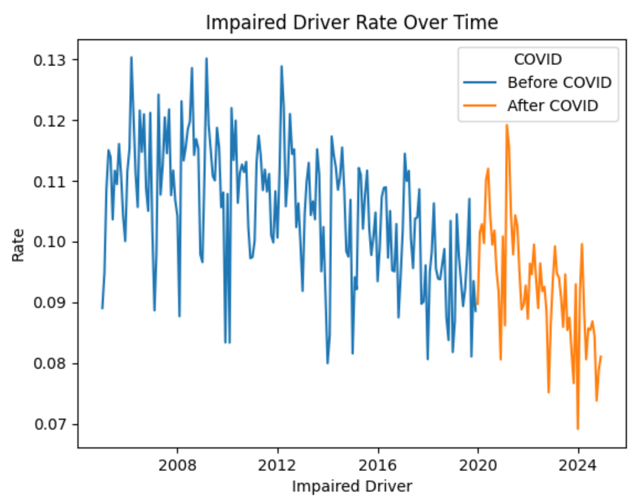
The drinking driving rate was in decline until COVID and then plateaued at the beginning of the pandemic.



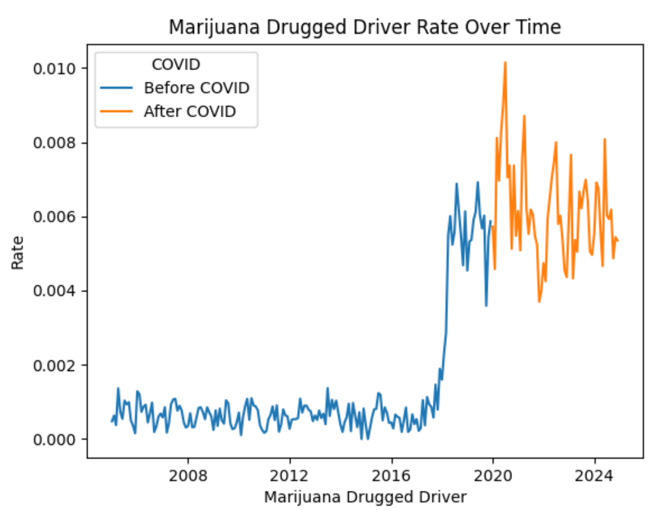
The drugged driver rate spiked at the start of COVID and then began to decline starting in 2021 or 2022.



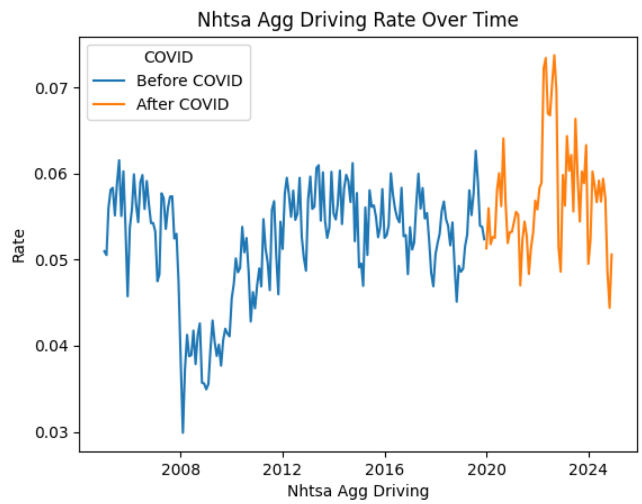
The hit and run rate spiked at the start of COVID and has been slowly declining since.



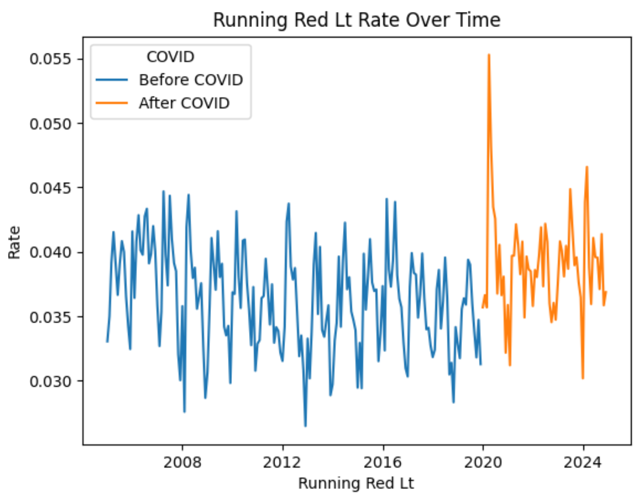
The impaired driver rate spiked a little bit during COVID in 2020 and into 2022 but has been declining since.



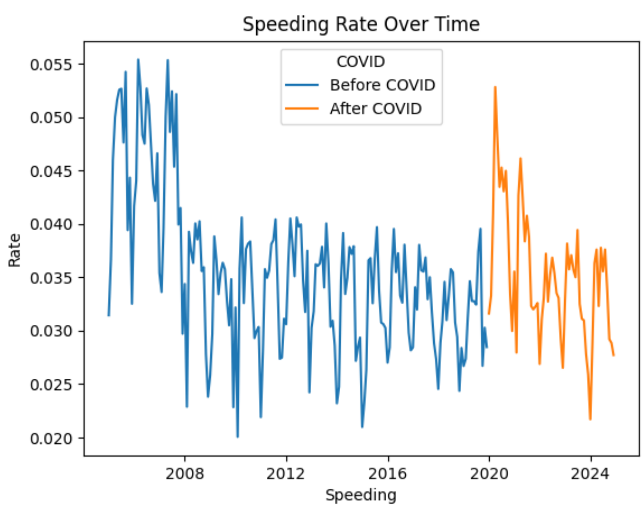
The marijuana drugged driver rate spike around 2018 before COVID and then spiked again during COVID. The trend has plateaued but remains high.



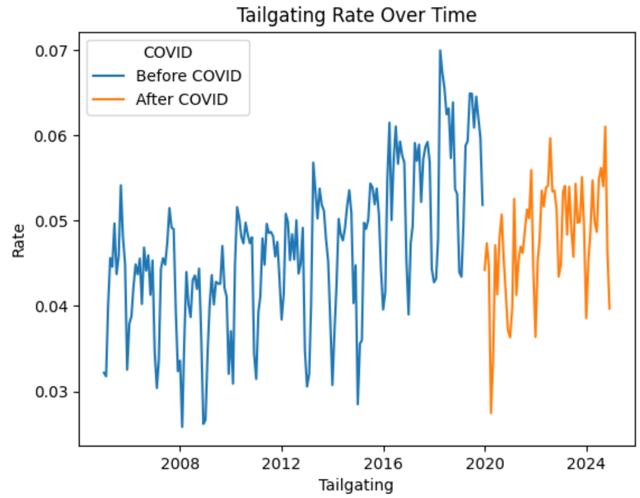
The NHTSA aggressive driving rate spiked in 2022 or 2023 and then declined to be closer to what it was from 2012 to 2018.



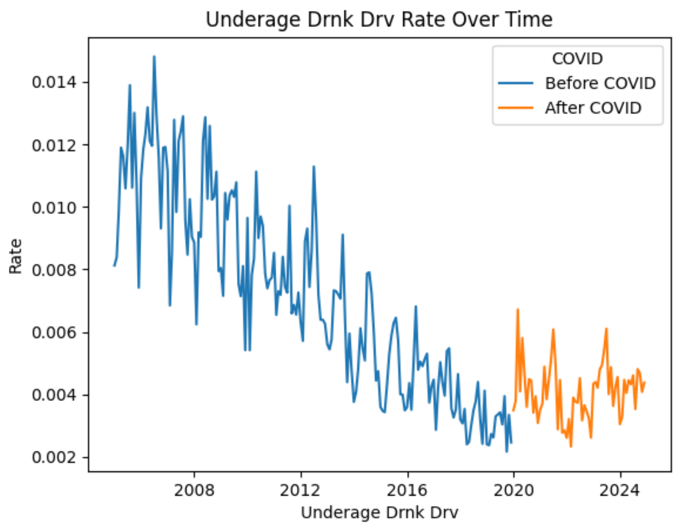
The rate of people running red lights spiked at the beginning of COVID and then declined. The trend remains a little higher than it was prior to COVID.



The speeding rate was high before 2008 but dropped significantly between 2008 and 2020 when it spiked again. The trend declined around 2022 to a rate closer to what it was from 2008 to 2020.



The tailgating rate dropped in 2020 and has been increasing since.



The underage drinking rate was declining before COVID. The trend plateaued in 2020 and remains steady.

Hypothesis Generation

It is surprising that specific aggressive driving behaviors, such as drugged driver rate and hit and run, spiked at the beginning of COVID while the overall aggressive driving rate did not appear to increase at the same time. This may be because aggressive drivers engage in more than one specific behavior, and engaging in more aggressive driving behaviors is what increased. This could result in a higher rate of crashes due to aggressive driving and may explain why people agree that aggressive driving behaviors increased with COVID. An increase in the number of behaviors drivers engage in could make their aggressiveness more noticeable and dangerous.

2.4 Data Quality Assessment

Error! Reference source not found. and ***Error! Reference source not found.*** shows the number an d percentage of missing values for each selected field in the CRASH and FLAGS datasets, respectively.

Completeness Analysis

The variables below have significant missingness. and will be removed from the dataset:

- ILLUMINATION has 2,448,908 (99.5%) missing values.
- WEATHER1 has 2,315,023 (94.1%) missing values.
- WEATHER2 has 2,161,825 (87.8%) missing values.

The following variables have 0 (0.0%) missing values:

- All FLAGS variables
- CRASH_MONTH, CRASH_YEAR, and DAY_OF_WEEK
- All DRIVER_COUNT_<age> variables
- COUNTY and URBAN_RURAL
- INTERSECT_TYPE
- LOCATION_TYPE
- ROAD_CONDITION

Accuracy Assessment

The FLAGS dataset contains some variables whose values correspond in some ways to the CRASH fields above.

The ILLUMINATION_DARK field from FLAGS, is 1 when ILLUMINATION is 'Dark streetlights', 'Dark - no streetlights', or 'Dark unknown roadway lighting.' However, ILLUMINATION_DARK is also 1 when ILLUMINATION is null for 817,191 records. The WEATHER1 and WEATHER2 fields were not used to derive ILLUMINATION_DARK because all values of weather are represented when ILLUMINATION_DARK is 1.

The INTERSECTION and NON_INTERSECTION field from FLAGS, corresponds with INTERSECT_TYPE where INTERSECTION is 1 for all values of INTERSECT_TYPE except 'Mid-block' and NON_INTERSECTION is 1 for 'Mid-block.'

The CROSS_MEDIAN, CURVE_DVR_ERROR, CURVED_ROAD, RAMP, RAMP_SEGMENT, RAMP_TERMINAL, ROUNDABOUT, OR SPEED_CHANGE_LANE fields from FLAGS do not appear to correspond with the LOCATION_TYPE field from CRASH.

The ICY_ROAD, SNOW_SLUSH_ROAD, and WET_ROAD fields from FLAGS correspond perfectly with ROAD_CONDITION values from CRASH. When ICY_ROAD is 1, ROAD_CONDITION is ''Ice/Frost'; when SNOW_SLUSH_ROAD is 1, ROAD_CONDITION is 'Snow' or 'Slush'; and when WET_ROAD is 1, ROAD_CONDITION is 'Wet' or 'Water (Standing or Moving).'

The URBAN and RURAL fields from FLAGS mostly correspond with URBAN_RURAL values from CRASH, which allows for 'Urban', 'Rural,' and 'Urbanized' values. The RURAL field is 1 for 818,102 values of 'Rural' and 6,220 for 'Urbanized' while the URBAN field is 1 for 1,636,862 values of 'Urbanized' and 0 for 9 values of 'Rural.' There are no instances of 'Urban,' and it is unclear what 'Urbanized' means. It could mean a small urban-like town in an otherwise rural area.

The DRIVER_<age> variables from FLAGS mostly correspond with the respective DRIVER_COUNT_<age> variables in CRASH. For each count variable in CRASH, there are a few values for which the count is 0 while the respective FLAGS variable is 1, however, these account for less than 0.00005% of records for each variable. Otherwise, the FLAGS variables are 1 when the count variables in CRASH are greater than 0.

Due to the nature of how these data were compiled and prepared for public consumption, it is not possible to verify the accuracy of the data. However, the primary variables chosen for this report have no missing values, and the correspondence between similar variables in CRASH and FLAGS is fairly consistent. The data appear to be accurate and can be used for analysis.

Consistency Evaluation

There are some inconsistencies between the FLAGS variables and their possible corresponding categorical variables in CRASH. There is no need to use similar variables from both datasets, or example, this project does not use both INTERSECT_TYPE from CRASH and INTERSECTION and NON_INTERSECTION from FLAGS. Temporal and geographic variables from CRASH are used for grouping and summarizing, and rates of occurrence of FLAGS variables are analyzed within these temporal or geographic categories. FLAGS variables that do not represent an aggressive driving behavior, such as DRIVER_<age>, are used to analyze covariance with aggressive driving.

CRN

There are no CRN values in other datasets that are not in CRASH. Every record in CRASH has a corresponding record in FLAGS.

Timelines and Currency

The data are current for the timeline of this project. The trend of various FLAGS categories is well established since it starts in 2005 so changes can be easily visualized and analyzed.

Relevance and Coverage

The data include crashes for the state of Pennsylvania. All counties are represented in the data.

Data Quality and Impact Assessment

The data are in condition for analysis. The variables ILLUMINATION, WEATHER1, and WEATHER2 that were originally chosen for this project have 99.5%, 94.1%, and 87.8% missing values, respectively. It would not make sense to impute that many values. All other variables chosen for this project have 0 missing values. The dataset is in good condition for analysis.

Appendix A: Data Dictionary for CRASH

Column	Description (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
COUNTY	County in PA where the crash occurred	int64	67	0 (0.0%)	[1, 67]
CRASH_MONTH	Month of the crash date	int64	12	0 (0.0%)	[1, 12]
CRASH_YEAR	Year of the crash	int64	20	0 (0.0%)	[2005, 2024]
CRN (PK)	Crash record number - unique identifier	int64	2461193	0 (0.0%)	[2005000003, 2025050430]
DAY_OF_WEEK	Day of week that the crash occurred	int64	7	0 (0.0%)	[1, 7]
DRIVER_COUNT_16YR	Number of 16-year-old drivers involved in the crash	int64	5	0 (0.0%)	[0, 4]
DRIVER_COUNT_17YR	Number of 17-year-old drivers involved in the crash	int64	7	0 (0.0%)	[0, 6]
DRIVER_COUNT_18YR	Number of 18-year-old drivers involved in the crash	int64	6	0 (0.0%)	[0, 5]
DRIVER_COUNT_19YR	Number of 19-year-old drivers involved in the crash	int64	5	0 (0.0%)	[0, 4]
DRIVER_COUNT_20YR	Number of 20-year-old drivers involved in the crash	int64	4	0 (0.0%)	[0, 3]
DRIVER_COUNT_50_64YR	Number of 50 to 64-year-old drivers involved in the crash	int64	16	0 (0.0%)	[0, 26]
DRIVER_COUNT_65_74YR	Number of 65 to 74-year-old drivers involved in the crash	int64	7	0 (0.0%)	[0, 7]
DRIVER_COUNT_75PLUS	Number of 75-year-old and older drivers involved in the crash	int64	5	0 (0.0%)	[0, 4]
ILLUMINATION †	Code that identifies the lighting at the time of the crash in terms time of day and existence of street lights	float64	7	2448908 (99.5%)	[1.0, 8.0]
INTERSECT_TYPE †	Code that identifies the type of intersection, if applicable, where the crash occurred	int64	14	0 (0.0%)	[0, 99]
LOCATION_TYPE †	Code that identifies the type of location where the crash occurred	int64	10	0 (0.0%)	[0, 99]
ROAD_CONDITION †	Code that identifies the condition of the road	int64	11	0 (0.0%)	[1, 99]

Column	Description (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
SECONDARY_CRASH †	Did a prior crash contribute in any way to this crash?	string	2	1884162 (76.55%)	<NA>, N, Y
URBAN_RURAL †	Code that identifies the location of the crash as either urban or rural	int64	2	0 (0.0%)	[1, 2]
WEATHER1 †	Code that identifies the primary weather condition at the time of the crash	float64	12	2315023 (94.06%)	[1.0, 99.0]
WEATHER2 †	Code that identifies the secondary weather condition at the time of the crash	float64	11	2161825 (87.84%)	[1.0, 98.0]

† See *Appendix C: Coded CRASH Variable Fields* and *Appendix D: Enumeration of Counties* for the definitions of enumerated values for each field.

Appendix B: Data Dictionary for FLAGS

Column	Definition (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
AGGRESSIVE_DRIVING	At least one aggressive driver action	int64	2	0 (0.0%)	[0, 1]
ALCOHOL_RELATED	At least one driver or pedestrian was reported or suspected of alcohol use	int64	2	0 (0.0%)	[0, 1]
CELL_PHONE	Driver was using a cell phone, either hand held or hands free	int64	2	0 (0.0%)	[0, 1]
CRN (PK, FK)	Crash record number - unique identifier	int64	2461193	0 (0.0%)	[2005000003, 2025050430]
CROSS_MEDIAN	At least one unit crossed a median	int64	2	0 (0.0%)	[0, 1]
CURVE_DVR_ERROR	At least one driver error in curve negotiation	int64	2	0 (0.0%)	[0, 1]
CURVED_ROAD	A curve in the road where the crash occurred	int64	2	0 (0.0%)	[0, 1]
DISTRACTED	At least one driver action occurred because of distraction	int64	2	0 (0.0%)	[0, 1]
DRINKING_DRIVER	At least one driver was drinking	int64	2	0 (0.0%)	[0, 1]
DRIVER_16YR	At least one driver was 16 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_17YR	At least one driver was 17 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_18YR	At least one driver was 18 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_19YR	At least one driver was 19 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_20YR	At least one driver was 20 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_50_64YR	At least one driver was 50 to 64 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_65_74YR	At least one driver was 65 to 74 years of age	int64	2	0 (0.0%)	[0, 1]
DRIVER_75PLUS	At least one driver was 75 years of age or older	int64	2	0 (0.0%)	[0, 1]

Column	Definition (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
DRUG_RELATED	A driver or non-motorist (cyclist or pedestrian) had a condition of drug use, was suspected of drug use by the police, or had a positive drug test indicating the presence of a controlled substance (this definition changed in May 2022)	int64	2	0 (0.0%)	[0, 1]
DRUGGED_DRIVER	Any driver had a condition of drug use, was suspected of drug use by the police, or had a positive drug test indicating the presence of a controlled substance (this definition changed in May 2022)	int64	2	0 (0.0%)	[0, 1]
FATIGUE_ASLEEP	At least one driver was fatigued or fell asleep	int64	2	0 (0.0%)	[0, 1]
HIT_RUN	At least one driver engaged in hit and run	int64	2	0 (0.0%)	[0, 1]
ICY_ROAD	Road was icy at the time of the crash	int64	2	0 (0.0%)	[0, 1]
ILLEGAL_DRUG_RELATED	At least one driver or pedestrian was reported or suspected of using illegal drugs	int64	2	0 (0.0%)	[0, 1]
ILLUMINATION_DARK	Crash site lighting was dark	int64	2	0 (0.0%)	[0, 1]
IMPAIRED_DRIVER	At least one driver was impaired by drugs or alcohol	int64	2	0 (0.0%)	[0, 1]
IMPAIRED_NONMOTORIST	At least one non-motorist was impaired by drugs or alcohol	int64	2	0 (0.0%)	[0, 1]
INTERSECTION	Crash occurred at an intersection	int64	2	0 (0.0%)	[0, 1]
LANE_DEPARTURE	At least one driver departed their lane of travel during the crash	int64	2	0 (0.0%)	[0, 1]
MARIJUANA_DRUGGED_DRIVER	At least one driver tested positive for marijuana	int64	2	0 (0.0%)	[0, 1]

Column	Definition (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
MARIJUANA_RELATED	At least one driver, pedestrian, or other non-motorist tested positive for marijuana	int64	2	0 (0.0%)	[0, 1]
MATURE_DRIVER	At least one driver was over the age of 65	int64	2	0 (0.0%)	[0, 1]
MC_DRINKING_DRIVER	At least one motorcycle driver was reported or suspected of alcohol use	int64	2	0 (0.0%)	[0, 1]
NHTSA_AGG_DRIVING	The crash meets the definition for aggressive driving established by NHTSA	int64	2	0 (0.0%)	[0, 1]
NON_INTERSECTION	Crash did not take place at an intersection	int64	2	0 (0.0%)	[0, 1]
OPIOID_RELATED	At least one motorcycle driver was reported or suspected of drug use and tested positive for opioids	int64	2	0 (0.0%)	[0, 1]
RAMP	Crash involved an interchange ramp	int64	2	0 (0.0%)	[0, 1]
RAMP_SEGMENT	Crash occurred on an interchange ramp between the beginning and end of the ramp	int64	2	0 (0.0%)	[0, 1]
RAMP_TERMINAL	Crash occurred at the end of an interchange ramp where a limited access highway meets a non-limited access highway	int64	2	0 (0.0%)	[0, 1]
ROUNDABOUT	Crash occurred at a modern roundabout intersection	int64	2	0 (0.0%)	[0, 1]
RUNNING_RED_LT	At least one driver ran a red light	int64	2	0 (0.0%)	[0, 1]
RUNNING_STOP_SIGN	At least one driver ran a stop sign	int64	2	0 (0.0%)	[0, 1]
RURAL	Crash occurred in a rural municipality	int64	2	0 (0.0%)	[0, 1]
SIGNALIZED_INT	Crash occurred at a signalized intersection	int64	2	0 (0.0%)	[0, 1]
SNOW_SLUSH_ROAD	Either snow or slush was on the road when the crash occurred	int64	2	0 (0.0%)	[0, 1]
SPEED_CHANGE_LANE	Crash occurred where an acceleration and deceleration lane was present on a limited access highway	int64	2	0 (0.0%)	[0, 1]

Column	Definition (PENNDOT, 2025)	Data Type	Number of Unique	Missing	Range/Unique
SPEEDING	At least one driver was speeding	int64	2	0 (0.0%)	[0, 1]
SPEEDING_RELATED	At least one driver was speeding, racing, or driving too fast for conditions	int64	2	0 (0.0%)	[0, 1]
STOP_CONTROLLED_INT	Crash occurred at a stop controlled intersection	int64	2	0 (0.0%)	[0, 1]
SUDDEN_DEER	Crash involved a deer in the roadway	int64	2	0 (0.0%)	[0, 1]
TAILGATING	At least one driver was tailgating or following too closely	int64	2	0 (0.0%)	[0, 1]
UNDERAGE_DRNK_DRV	At least one driver was underage and drinking	int64	2	0 (0.0%)	[0, 1]
UNLICENSED	At least one driver was unlicensed	int64	2	0 (0.0%)	[0, 1]
UNSIGNALIZED_INT	Crash occurred at an unsignalized intersection	int64	2	0 (0.0%)	[0, 1]
URBAN	Crash occurred in a n urban municipality	int64	2	0 (0.0%)	[0, 1]
WET_ROAD	Road was wet at the time of the crash	int64	2	0 (0.0%)	[0, 1]
YOUNG_DRIVER	At least one driver action occurred because of distraction	int64	2	0 (0.0%)	[0, 1]

Appendix C: Coded CRASH Variable Fields

COUNTY

See **Error! Not a valid bookmark self-reference.**

DAY_OF_WEEK

1: 'Sunday'
2: 'Monday'
3: 'Tuesday'
4: 'Wednesday'
5: 'Thursday'
6: 'Friday'
7: 'Saturday'
9: ‘Unknown

ILLUMINATION

1: 'Daylight'
2: 'Dark - no streetlights'
3: 'Dark streetlights'
4: 'Dusk'
5: 'Dawn'
6: 'Dark unknown roadway lighting'
8: 'Other'
9: 'Unknown'

INTERSECT_TYPE

0: 'Mid-block'
1: 'Four-way intersection'
2: '“T” intersection'
3: '“Y” intersection'
4: 'Traffic Circle/Roundabout (EXPIRED 1/1/18)'
5: 'Multi-leg intersection'
6: 'Ramp End'
7: 'Ramp Begin'
8: 'Crossover'
9: 'Railroad crossing'
10: 'Other'
11: '“L” Intersection'
12: 'Traffic Circle'
13: 'Roundabout'
99: 'Unknown'

LOCATION_TYPE

0: 'Not applicable'
1: 'Underpass'
2: 'Ramp'
3: 'Bridge'
4: 'Tunnel'
5: 'Toll Booth'
6: 'Cross over related'
7: 'Driveway or Parking Lot'
8: 'Ramp and bridge'
99: 'Unknown'

ROAD_CONDITION

1: 'Dry'
2: 'Ice/Frost'
3: 'Mud, Dirt, Gravel'
4: 'Oil'
5: 'Sand'
6: 'Slush'
7: 'Snow'
8: 'Water (Standing or Moving)'
9: 'Wet'
22: 'Mud, Sand, Dirt, Oil (Expired 1-1-20)'
98: 'Other'
99: 'Unknown'

SECONDARY_CRASH

Y: 'Yes'
N: 'No'
<NA>: missing

URBAN_RURAL

1: 'Rural'
2: 'Urbanized'
3: 'Urban'

WEATHER1/WEATHER2

1: 'Blowing Sand, Soil, Dirt'
2: 'Blowing Snow'
3: 'Clear'
4: 'Cloudy'
5: 'Fog, Smog, Smoke'
6: 'Freezing Rain or Freezing Drizzle'
7: 'Rain'
8: 'Severe Crosswinds'
9: 'Sleet or Hail'
10: 'Snow'
98: 'Other'
99: 'Unknown' ‡

‡ WEATHER1 only

Appendix D: Enumeration of Counties

County	Enumeration
ADAMS	1
ALLEGHENY	2
ARMSTRONG	3
BEAVER	4
BEDFORD	5
BERKS	6
BLAIR	7
BRADFORD	8
BUCKS	9
BUTLER	10
CAMBRIA	11
CAMERON	12
CARBON	13
CENTRE	14
CHESTER	15
CLARION	16
CLEARFIELD	17
CLINTON	18
COLUMBIA	19
CRAWFORD	20
CUMBERLAND	21
DAUPHIN	22
DELAWARE	23
ELK	24
ERIE	25
FAYETTE	26
FOREST	27
FRANKLIN	28
FULTON	29
GREENE	30
HUNTINGDON	31
INDIANA	32
JEFFERSON	33
JUNIATA	34
LACKAWANNA	35

County	Enumeration
LANCASTER	36
LAWRENCE	37
LEBANON	38
LEHIGH	39
LUZERNE	40
LYCOMING	41
MCKEAN	42
MERCER	43
MIFFLIN	44
MONROE	45
MONTGOMERY	46
MONTOUR	47
NORTHAMPTON	48
NORTHUMBERLAND	49
PERRY	50
PIKE	51
POTTER	52
SCHUYLKILL	53
SNYDER	54
SOMERSET	55
SULLIVAN	56
SUSQUEHANNA	57
TIOGA	58
UNION	59
VENANGO	60
WARREN	61
WASHINGTON	62
WAYNE	63
WESTMORELAND	64
WYOMING	65
YORK	66
PHILADELPHIA	67

Appendix E: File Information

Each file from the *PennDOT Crash Data Download* portal consists of tabular data in a comma separated values (CSV) file. The table below displays the year, file path, file size, number of rows, and number of columns for each dataset.

Year	File Path	File Size (MB)	Number of Rows	Number of Columns
2005	data/raw/Statewide_2005/COMMVEH_2005.csv	1.52	8415	32
2005	data/raw/Statewide_2005/CRASH_2005.csv	54.15	134,261	99
2005	data/raw/Statewide_2005/CYCLE_2005.csv	0.48	5,788	21
2005	data/raw/Statewide_2005/FLAGS_2005.csv	68.38	134,261	130
2005	data/raw/Statewide_2005/PERSON_2005.csv	32.32	323,241	23
2005	data/raw/Statewide_2005/ROADWAY_2005.csv	15.76	204,591	13
2005	data/raw/Statewide_2005/TRAILVEH_2005.csv	0.4	9,801	8
2005	data/raw/Statewide_2005/VEHICLE_2005.csv	43.05	236,909	41
2006	data/raw/Statewide_2006/COMMVEH_2006.csv	1.35	7,450	32
2006	data/raw/Statewide_2006/CRASH_2006.csv	52.5	129,253	99
2006	data/raw/Statewide_2006/CYCLE_2006.csv	0.46	5,571	21
2006	data/raw/Statewide_2006/FLAGS_2006.csv	65.83	129,253	130
2006	data/raw/Statewide_2006/PERSON_2006.csv	31.17	311,602	23
2006	data/raw/Statewide_2006/ROADWAY_2006.csv	15.47	200,835	13
2006	data/raw/Statewide_2006/TRAILVEH_2006.csv	0.27	6,140	8
2006	data/raw/Statewide_2006/VEHICLE_2006.csv	41.87	229,365	41
2007	data/raw/Statewide_2007/COMMVEH_2007.csv	1.53	8,448	32
2007	data/raw/Statewide_2007/CRASH_2007.csv	54.72	132,152	99
2007	data/raw/Statewide_2007/CYCLE_2007.csv	0.48	5,811	21
2007	data/raw/Statewide_2007/FLAGS_2007.csv	67.3	132,152	130
2007	data/raw/Statewide_2007/PERSON_2007.csv	31.4	313,795	23
2007	data/raw/Statewide_2007/ROADWAY_2007.csv	15.31	198,628	13
2007	data/raw/Statewide_2007/TRAILVEH_2007.csv	0.3	6,989	8
2007	data/raw/Statewide_2007/VEHICLE_2007.csv	42.24	231,408	41
2008	data/raw/Statewide_2008/COMMVEH_2008.csv	1.36	7,512	32
2008	data/raw/Statewide_2008/CRASH_2008.csv	51.96	126,184	99
2008	data/raw/Statewide_2008/CYCLE_2008.csv	0.45	5,254	21
2008	data/raw/Statewide_2008/FLAGS_2008.csv	64.26	126,184	130
2008	data/raw/Statewide_2008/PERSON_2008.csv	29.49	293,312	23
2008	data/raw/Statewide_2008/ROADWAY_2008.csv	14.54	187,973	13
2008	data/raw/Statewide_2008/TRAILVEH_2008.csv	0.26	5,651	8

Year	File Path	File Size (MB)	Number of Rows	Number of Columns
2008	data/raw/Statewide_2008/VEHICLE_2008.csv	39.81	218,713	41
2009	data/raw/Statewide_2009/COMMVEH_2009.csv	1.2	6,620	32
2009	data/raw/Statewide_2009/CRASH_2009.csv	50.13	121,794	99
2009	data/raw/Statewide_2009/CYCLE_2009.csv	0.41	4,854	21
2009	data/raw/Statewide_2009/FLAGS_2009.csv	62.03	121,794	130
2009	data/raw/Statewide_2009/PERSON_2009.csv	28.93	287,324	23
2009	data/raw/Statewide_2009/ROADWAY_2009.csv	14.5	186,702	13
2009	data/raw/Statewide_2009/TRAILVEH_2009.csv	0.21	4,520	8
2009	data/raw/Statewide_2009/VEHICLE_2009.csv	38.82	213,217	41
2010	data/raw/Statewide_2010/COMMVEH_2010.csv	1.33	7,307	32
2010	data/raw/Statewide_2010/CRASH_2010.csv	50.08	121,612	99
2010	data/raw/Statewide_2010/CYCLE_2010.csv	0.44	5,233	21
2010	data/raw/Statewide_2010/FLAGS_2010.csv	61.93	121,612	130
2010	data/raw/Statewide_2010/PERSON_2010.csv	29.14	289,421	23
2010	data/raw/Statewide_2010/ROADWAY_2010.csv	14.66	188,530	13
2010	data/raw/Statewide_2010/TRAILVEH_2010.csv	0.24	5,001	8
2010	data/raw/Statewide_2010/VEHICLE_2010.csv	39.29	215,820	41
2011	data/raw/Statewide_2011/COMMVEH_2011.csv	1.43	7,846	32
2011	data/raw/Statewide_2011/CRASH_2011.csv	51.79	125,616	99
2011	data/raw/Statewide_2011/CYCLE_2011.csv	0.39	4,674	21
2011	data/raw/Statewide_2011/FLAGS_2011.csv	63.97	125,616	130
2011	data/raw/Statewide_2011/PERSON_2011.csv	30.6	294,523	23
2011	data/raw/Statewide_2011/ROADWAY_2011.csv	15.2	195,679	13
2011	data/raw/Statewide_2011/TRAILVEH_2011.csv	0.26	5,402	8
2011	data/raw/Statewide_2011/VEHICLE_2011.csv	40.27	221,514	41
2012	data/raw/Statewide_2012/COMMVEH_2012.csv	1.31	7,220	32
2012	data/raw/Statewide_2012/CRASH_2012.csv	51.83	124,501	99
2012	data/raw/Statewide_2012/CYCLE_2012.csv	0.43	5,139	21
2012	data/raw/Statewide_2012/FLAGS_2012.csv	63.41	124,501	130
2012	data/raw/Statewide_2012/PERSON_2012.csv	30.26	291,142	23
2012	data/raw/Statewide_2012/ROADWAY_2012.csv	15.21	195,822	13
2012	data/raw/Statewide_2012/TRAILVEH_2012.csv	0.24	5,154	8
2012	data/raw/Statewide_2012/VEHICLE_2012.csv	40.03	220,251	41
2013	data/raw/Statewide_2013/COMMVEH_2013.csv	1.36	7,517	32
2013	data/raw/Statewide_2013/CRASH_2013.csv	51.42	124,366	99
2013	data/raw/Statewide_2013/CYCLE_2013.csv	0.39	4,732	21
2013	data/raw/Statewide_2013/FLAGS_2013.csv	63.34	124,366	130
2013	data/raw/Statewide_2013/PERSON_2013.csv	28.92	287,126	23

Year	File Path	File Size (MB)	Number of Rows	Number of Columns
2013	data/raw/Statewide_2013/ROADWAY_2013.csv	15.16	194,817	13
2013	data/raw/Statewide_2013/TRAILVEH_2013.csv	0.25	5,284	8
2013	data/raw/Statewide_2013/VEHICLE_2013.csv	39.79	218,867	41
2014	data/raw/Statewide_2014/COMMVEH_2014.csv	1.46	8,070	32
2014	data/raw/Statewide_2014/CRASH_2014.csv	50.24	121,547	99
2014	data/raw/Statewide_2014/CYCLE_2014.csv	0.38	4,590	21
2014	data/raw/Statewide_2014/FLAGS_2014.csv	61.9	121,547	130
2014	data/raw/Statewide_2014/PERSON_2014.csv	28.19	279,913	23
2014	data/raw/Statewide_2014/ROADWAY_2014.csv	14.61	188,456	13
2014	data/raw/Statewide_2014/TRAILVEH_2014.csv	0.27	5,608	8
2014	data/raw/Statewide_2014/VEHICLE_2014.csv	39.07	215,299	41
2015	data/raw/Statewide_2015/COMMVEH_2015.csv	1.55	8,548	32
2015	data/raw/Statewide_2015/CRASH_2015.csv	52.69	127,470	99
2015	data/raw/Statewide_2015/CYCLE_2015.csv	0.38	4,532	21
2015	data/raw/Statewide_2015/FLAGS_2015.csv	64.92	127,470	130
2015	data/raw/Statewide_2015/PERSON_2015.csv	29.58	293,690	23
2015	data/raw/Statewide_2015/ROADWAY_2015.csv	15.3	197,547	13
2015	data/raw/Statewide_2015/TRAILVEH_2015.csv	0.29	6,099	8
2015	data/raw/Statewide_2015/VEHICLE_2015.csv	41.14	226,713	41
2016	data/raw/Statewide_2016/COMMVEH_2016.csv	1.52	8,411	32
2016	data/raw/Statewide_2016/CRASH_2016.csv	53.44	129,607	99
2016	data/raw/Statewide_2016/CYCLE_2016.csv	0.39	4,653	21
2016	data/raw/Statewide_2016/FLAGS_2016.csv	66.01	129,607	130
2016	data/raw/Statewide_2016/PERSON_2016.csv	30.33	300,497	23
2016	data/raw/Statewide_2016/ROADWAY_2016.csv	15.77	203,624	13
2016	data/raw/Statewide_2016/TRAILVEH_2016.csv	0.29	6,051	8
2016	data/raw/Statewide_2016/VEHICLE_2016.csv	42.36	233,443	41
2017	data/raw/Statewide_2017/COMMVEH_2017.csv	1.54	8,516	32
2017	data/raw/Statewide_2017/CRASH_2017.csv	53.34	128,441	99
2017	data/raw/Statewide_2017/CYCLE_2017.csv	0.35	4,226	21
2017	data/raw/Statewide_2017/FLAGS_2017.csv	65.39	128,441	130
2017	data/raw/Statewide_2017/PERSON_2017.csv	30.12	294,820	23
2017	data/raw/Statewide_2017/ROADWAY_2017.csv	15.75	201,444	13
2017	data/raw/Statewide_2017/TRAILVEH_2017.csv	0.29	6,107	8
2017	data/raw/Statewide_2017/VEHICLE_2017.csv	42.05	231,152	41
2018	data/raw/Statewide_2018/COMMVEH_2018.csv	1.65	9,135	32
2018	data/raw/Statewide_2018/CRASH_2018.csv	53.56	128,541	99
2018	data/raw/Statewide_2018/CYCLE_2018.csv	0.29	3,435	21

Year	File Path	File Size (MB)	Number of Rows	Number of Columns
2018	data/raw/Statewide_2018/FLAGS_2018.csv	65.45	128,541	130
2018	data/raw/Statewide_2018/PERSON_2018.csv	30.66	291,112	23
2018	data/raw/Statewide_2018/ROADWAY_2018.csv	15.7	200,721	13
2018	data/raw/Statewide_2018/TRAILVEH_2018.csv	0.3	6,347	8
2018	data/raw/Statewide_2018/VEHICLE_2018.csv	41.79	229,674	41
2019	data/raw/Statewide_2019/COMMVEH_2019.csv	1.55	8,578	32
2019	data/raw/Statewide_2019/CRASH_2019.csv	52.12	125,452	99
2019	data/raw/Statewide_2019/CYCLE_2019.csv	0.33	3,902	21
2019	data/raw/Statewide_2019/FLAGS_2019.csv	63.89	125,452	130
2019	data/raw/Statewide_2019/PERSON_2019.csv	29.62	284,241	23
2019	data/raw/Statewide_2019/ROADWAY_2019.csv	15.45	197,676	13
2019	data/raw/Statewide_2019/TRAILVEH_2019.csv	0.29	6,119	8
2019	data/raw/Statewide_2019/VEHICLE_2019.csv	41.01	225,732	41
2020	data/raw/Statewide_2020/COMMVEH_2020.csv	1.36	7,498	32
2020	data/raw/Statewide_2020/CRASH_2020.csv	44.25	104,600	99
2020	data/raw/Statewide_2020/CYCLE_2020.csv	0.36	4,185	21
2020	data/raw/Statewide_2020/FLAGS_2020.csv	53.27	104,600	130
2020	data/raw/Statewide_2020/PERSON_2020.csv	23.51	224,455	23
2020	data/raw/Statewide_2020/ROADWAY_2020.csv	12.55	162,874	13
2020	data/raw/Statewide_2020/TRAILVEH_2020.csv	0.26	5,376	8
2020	data/raw/Statewide_2020/VEHICLE_2020.csv	33.45	184,240	41
2021	data/raw/Statewide_2021/COMMVEH_2021.csv	1.58	8,720	32
2021	data/raw/Statewide_2021/CRASH_2021.csv	49.69	118,100	99
2021	data/raw/Statewide_2021/CYCLE_2021.csv	0.37	4,344	21
2021	data/raw/Statewide_2021/FLAGS_2021.csv	60.15	118,100	130
2021	data/raw/Statewide_2021/PERSON_2021.csv	26.78	258,421	23
2021	data/raw/Statewide_2021/ROADWAY_2021.csv	14.28	185,199	13
2021	data/raw/Statewide_2021/TRAILVEH_2021.csv	0.29	5,940	8
2021	data/raw/Statewide_2021/VEHICLE_2021.csv	38.44	212,073	41
2022	data/raw/Statewide_2022/COMMVEH_2022.csv	1.76	9,590	32
2022	data/raw/Statewide_2022/CRASH_2022.csv	49.24	116,147	99
2022	data/raw/Statewide_2022/CYCLE_2022.csv	0.36	4,210	21
2022	data/raw/Statewide_2022/FLAGS_2022.csv	59.15	116,147	130
2022	data/raw/Statewide_2022/PERSON_2022.csv	26.67	254,097	23
2022	data/raw/Statewide_2022/ROADWAY_2022.csv	14.09	182,181	13
2022	data/raw/Statewide_2022/TRAILVEH_2022.csv	0.34	5,872	8
2022	data/raw/Statewide_2022/VEHICLE_2022.csv	37.38	206,698	41
2023	data/raw/Statewide_2023/COMMVEH_2023.csv	1.58	8,585	32

Year	File Path	File Size (MB)	Number of Rows	Number of Columns
2023	data/raw/Statewide_2023/CRASH_2023.csv	46.89	110,736	99
2023	data/raw/Statewide_2023/CYCLE_2023.csv	0.36	4,322	21
2023	data/raw/Statewide_2023/FLAGS_2023.csv	56.4	110,736	130
2023	data/raw/Statewide_2023/PERSON_2023.csv	26.36	245,499	23
2023	data/raw/Statewide_2023/ROADWAY_2023.csv	13.51	174,277	13
2023	data/raw/Statewide_2023/TRAILVEH_2023.csv	0.3	5,072	8
2023	data/raw/Statewide_2023/VEHICLE_2023.csv	35.84	198,689	41
2024	data/raw/Statewide_2024/COMMVEH_2024.csv	1.6	8,716	32
2024	data/raw/Statewide_2024/CRASH_2024.csv	47.11	110,813	99
2024	data/raw/Statewide_2024/CYCLE_2024.csv	0.3	3,426	21
2024	data/raw/Statewide_2024/FLAGS_2024.csv	56.43	110,813	130
2024	data/raw/Statewide_2024/PERSON_2024.csv	26.19	244,173	23
2024	data/raw/Statewide_2024/ROADWAY_2024.csv	13.57	174,249	13
2024	data/raw/Statewide_2024/TRAILVEH_2024.csv	0.3	5,179	8
2024	data/raw/Statewide_2024/VEHICLE_2024.csv	35.63	197,506	41

Appendix F: Columns and Data Types

The table below displays the dataset name and data type for each column in each of the eight categories of data, e.g., COMMVEH, CRASH, CYCLE, etc. The specified data types are what Python assigned when reading the data.
Not all columns will be selected for use in this project. Columns chosen for this project that are data type 'object' will need to be converted to string, float, or integer after evaluating which type is appropriate for each one.

Dataset	Column	Data Type
COMMVEH	CRN	int64
COMMVEH	AXLE_CNT	float64
COMMVEH	CARGO_BD_TYPE	float64
COMMVEH	CARRIER_ADDR_1	str
COMMVEH	CARRIER_ADDR_2	str
COMMVEH	CARRIER_ADDR_CITY	str
COMMVEH	CARRIER_ADDR_STATE	str
COMMVEH	CARRIER_ADDR_ZIP	str
COMMVEH	CARRIER_NM	str
COMMVEH	CARRIER_TEL	object
COMMVEH	GVWR	object
COMMVEH	HAZMAT_CD1	float64
COMMVEH	HAZMAT_CD2	float64
COMMVEH	HAZMAT_CD3	float64
COMMVEH	HAZMAT_CD4	float64
COMMVEH	HAZMAT_IND	str
COMMVEH	HAZMAT_REL_IND1	float64
COMMVEH	HAZMAT_REL_IND2	float64
COMMVEH	HAZMAT_REL_IND3	float64
COMMVEH	HAZMAT_REL_IND4	float64
COMMVEH	ICC_NUM	object
COMMVEH	OSIZE_LOAD_IND	str
COMMVEH	PERMITTED	float64
COMMVEH	PUC_NUM	str
COMMVEH	SPECIAL_SIZING1	float64
COMMVEH	SPECIAL_SIZING2	float64
COMMVEH	SPECIAL_SIZING3	float64
COMMVEH	SPECIAL_SIZING4	float64
COMMVEH	TYPE_OF_CARRIER	float64
COMMVEH	UNIT_NUM	int64
COMMVEH	USDOT_NUM	str

Dataset	Column	Data Type
COMMVEH	VEH_CONFIG_CD	float64
CRASH	CRN	int64
CRASH	ARRIVAL_TM	float64
CRASH	AUTOMOBILE_COUNT	int64
CRASH	BELTED_DEATH_COUNT	int64
CRASH	BELTED_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	BICYCLE_COUNT	int64
CRASH	BICYCLE_DEATH_COUNT	int64
CRASH	BICYCLE_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	BUS_COUNT	int64
CRASH	CHLDPAS_DEATH_COUNT	int64
CRASH	CHLDPAS_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	COLLISION_TYPE	int64
CRASH	COMM_VEH_COUNT	int64
CRASH	CONS_ZONE_SPD_LIM	float64
CRASH	COUNTY	int64
CRASH	CRASH_MONTH	int64
CRASH	CRASH_YEAR	int64
CRASH	DAY_OF_WEEK	int64
CRASH	DEC_LATITUDE	float64
CRASH	DEC_LONGITUDE	float64
CRASH	DISPATCH_TM	float64
CRASH	DISTRICT	int64
CRASH	DRIVER_COUNT_16YR	int64
CRASH	DRIVER_COUNT_17YR	int64
CRASH	DRIVER_COUNT_18YR	int64
CRASH	DRIVER_COUNT_19YR	int64
CRASH	DRIVER_COUNT_20YR	int64
CRASH	DRIVER_COUNT_50_64YR	int64
CRASH	DRIVER_COUNT_65_74YR	int64
CRASH	DRIVER_COUNT_75PLUS	int64
CRASH	EST_HRS_CLOSED	float64
CRASH	FATAL_COUNT	int64
CRASH	HEAVY_TRUCK_COUNT	int64
CRASH	HORSE_BUGGY_COUNT	int64
CRASH	HOUR_OF_DAY	float64
CRASH	ILLUMINATION	float64
CRASH	INJURY_COUNT	int64

Dataset	Column	Data Type
CRASH	INTERSECTION_RELATED	object
CRASH	INTERSECT_TYPE	int64
CRASH	LANE_CLOSED	float64
CRASH	LATITUDE	str
CRASH	LN_CLOSE_DIR	float64
CRASH	LOCATION_TYPE	int64
CRASH	LONGITUDE	str
CRASH	MAX_SEVERITY_LEVEL	int64
CRASH	MCYCLE_DEATH_COUNT	int64
CRASH	MCYCLE_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	MOTORCYCLE_COUNT	int64
CRASH	MUNICIPALITY	int64
CRASH	NONMOTR_COUNT	int64
CRASH	NONMOTR_DEATH_COUNT	int64
CRASH	NONMOTR_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	NTFY_HIWY_MAINT	str
CRASH	PED_COUNT	int64
CRASH	PED_DEATH_COUNT	int64
CRASH	PED_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	PERSON_COUNT	int64
CRASH	POLICE_AGCY	str
CRASH	POSSIBLE_INJ_COUNT	int64
CRASH	RDWY_SURF_TYPE_CD	float64
CRASH	RELATION_TO_ROAD	float64
CRASH	ROADWAY_CLEARED	float64
CRASH	ROAD_CONDITION	int64
CRASH	SCH_BUS_IND	str
CRASH	SCH_ZONE_IND	str
CRASH	SECONDARY_CRASH	object
CRASH	SMALL_TRUCK_COUNT	int64
CRASH	SPEC_JURIS_CD	float64
CRASH	SUSP_MINOR_INJ_COUNT	int64
CRASH	SUSP_SERIOUS_INJ_COUNT	int64
CRASH	SUV_COUNT	int64
CRASH	TCD_FUNC_CD	float64
CRASH	TCD_TYPE	int64
CRASH	TFC_DETOUR_IND	str
CRASH	TIME_OF_DAY	float64

Dataset	Column	Data Type
CRASH	TOTAL_UNITS	int64
CRASH	TOT_INJ_COUNT	int64
CRASH	UNBELTED_OCC_COUNT	int64
CRASH	UNB_DEATH_COUNT	int64
CRASH	UNB_SUSP_SERIOUS_INJ_COUNT	int64
CRASH	UNK_INJ_DEG_COUNT	int64
CRASH	UNK_INJ_PER_COUNT	int64
CRASH	URBAN_RURAL	int64
CRASH	VAN_COUNT	int64
CRASH	VEHICLE_COUNT	int64
CRASH	WEATHER1	float64
CRASH	WEATHER2	float64
CRASH	WORKERS_PRES	str
CRASH	WORK_ZONE_IND	str
CRASH	WORK_ZONE_LOC	float64
CRASH	WORK_ZONE_TYPE	float64
CRASH	WZ_CLOSE_DETOUR	str
CRASH	WZ_FLAGGER	str
CRASH	WZ_LAW_OFFCR_IND	str
CRASH	WZ_LN_CLOSURE	str
CRASH	WZ_MOVING	str
CRASH	WZ_OTHER	str
CRASH	WZ_SHLDER_MDN	str
CRASH	WZ_WORKERS_INJ_KILLED	object
CYCLE	CRN	int64
CYCLE	MC_BAG_IND	str
CYCLE	MC_DVR_BOOTS_IND	str
CYCLE	MC_DVR_EDC_IND	str
CYCLE	MC_DVR_EYEPRT_IND	str
CYCLE	MC_DVR_HLMTDOT_IND	str
CYCLE	MC_DVR_HLMTON_IND	str
CYCLE	MC_DVR_HLMT_TYPE	float64
CYCLE	MC_DVR_LNGPNTS_IND	str
CYCLE	MC_DVR_LNGSLV_IND	str
CYCLE	MC_ENGINE_SIZE	object
CYCLE	MC_PASSNGR_IND	str
CYCLE	MC_PAS_BOOTS_IND	str
CYCLE	MC_PAS_EYEPRT_IND	str

Dataset	Column	Data Type
CYCLE	MC_PAS_HLMTDOT_IND	str
CYCLE	MC_PAS_HLMTON_IND	str
CYCLE	MC_PAS_HLMT_TYPE	float64
CYCLE	MC_PAS_LNGPNTS_IND	str
CYCLE	MC_PAS_LNGSLV_IND	str
CYCLE	MC_TRAIL_IND	str
CYCLE	UNIT_NUM	int64
FLAGS	CRN	int64
FLAGS	AGGRESSIVE_DRIVING	int64
FLAGS	ALCOHOL_RELATED	int64
FLAGS	ANGLE_CRASH	int64
FLAGS	ATV	int64
FLAGS	ATV_ROUTE	int64
FLAGS	BACKUP_CONGESTION	int64
FLAGS	BACKUP_NONRECURRING	int64
FLAGS	BACKUP_PRIOR	int64
FLAGS	BICYCLE	int64
FLAGS	CELL_PHONE	int64
FLAGS	CHILD_PASSENGER	int64
FLAGS	COMM_VEHICLE	int64
FLAGS	CORE_NETWORK	int64
FLAGS	CROSS_MEDIAN	int64
FLAGS	CURVED_ROAD	int64
FLAGS	CURVE_DVR_ERROR	int64
FLAGS	DEER_RELATED	int64
FLAGS	DISTRACTED	int64
FLAGS	DRINKING_DRIVER	int64
FLAGS	DRIVER_16YR	int64
FLAGS	DRIVER_17YR	int64
FLAGS	DRIVER_18YR	int64
FLAGS	DRIVER_19YR	int64
FLAGS	DRIVER_20YR	int64
FLAGS	DRIVER_50_64YR	int64
FLAGS	DRIVER_65_74YR	int64
FLAGS	DRIVER_75PLUS	int64
FLAGS	DRUGGED_DRIVER	int64
FLAGS	DRUG_RELATED	int64
FLAGS	FATAL	int64

Dataset	Column	Data Type
FLAGS	FATAL_OR_SUSP_SERIOUS_INJ	int64
FLAGS	FATIGUE_ASLEEP	int64
FLAGS	FEDERAL_AID_ROUTE	int64
FLAGS	FIRE_IN_VEHICLE	int64
FLAGS	HAZARDOUS_TRUCK	int64
FLAGS	HIT_BARRIER	int64
FLAGS	HIT_BRIDGE	int64
FLAGS	HIT_DEER	int64
FLAGS	HIT_EMBANKMENT	int64
FLAGS	HIT_FIXED_OBJECT	int64
FLAGS	HIT_GDRAIL	int64
FLAGS	HIT_GDRAIL_END	int64
FLAGS	HIT_PARKED_VEHICLE	int64
FLAGS	HIT_POLE	int64
FLAGS	HIT_ROADWAY_EQUIPMENT	int64
FLAGS	HIT_RUN	int64
FLAGS	HIT_TEMP_CONSTRUCTION_BARRIER	int64
FLAGS	HIT_TRAFFIC_ISLAND	int64
FLAGS	HIT_TREE_SHRUB	int64
FLAGS	HIT_UTILITY_POLE	int64
FLAGS	HORSE_BUGGY	float64
FLAGS	HO_OPPDIR_SDSWP	int64
FLAGS	HVY_TRUCK_RELATED	int64
FLAGS	ICY_ROAD	int64
FLAGS	ILLEGAL_DRUG_RELATED	int64
FLAGS	ILLUMINATION_DARK	int64
FLAGS	IMPAIRED_DRIVER	int64
FLAGS	IMPAIRED_NONMOTORIST	int64
FLAGS	INJURY	int64
FLAGS	INJURY_OR_FATAL	int64
FLAGS	INTERSECTION	int64
FLAGS	INTERSECTION_RELATED	int64
FLAGS	INTERSTATE	int64
FLAGS	LANE_DEPARTURE	int64
FLAGS	LEFT_TURN	int64
FLAGS	LIMIT_65MPH	int64
FLAGS	LIMIT_70MPH	float64
FLAGS	LOCAL_ROAD	int64

Dataset	Column	Data Type
FLAGS	LOCAL_ROAD_ONLY	int64
FLAGS	MARIJUANA_DRUGGED_DRIVER	int64
FLAGS	MARIJUANA_RELATED	int64
FLAGS	MATURE_DRIVER	int64
FLAGS	MC_DRINKING_DRIVER	int64
FLAGS	MOTORCYCLE	int64
FLAGS	MULTIPLE_VEHICLE	int64
FLAGS	NHTSA_AGG_DRIVING	int64
FLAGS	NON_INTERSECTION	int64
FLAGS	NO_CLEARANCE	int64
FLAGS	OPIOID_RELATED	int64
FLAGS	OTHER_FREEWAY_EXPRESSWAY	int64
FLAGS	OVERTURNED	int64
FLAGS	PEDESTRIAN	int64
FLAGS	PHANTOM_VEHICLE	int64
FLAGS	POSSIBLE_INJURY	int64
FLAGS	PROPERTY_DAMAGE_ONLY	int64
FLAGS	PSP_REPORTED	int64
FLAGS	RAMP	int64
FLAGS	RAMP_SEGMENT	int64
FLAGS	RAMP_TERMINAL	int64
FLAGS	REAR_END	int64
FLAGS	ROUNDABOUT	int64
FLAGS	RUNNING_RED_LT	int64
FLAGS	RUNNING_STOP_SIGN	int64
FLAGS	RURAL	int64
FLAGS	SCHOOL_BUS	int64
FLAGS	SCHOOL_BUS_RELATED	int64
FLAGS	SCHOOL_BUS_UNIT	int64
FLAGS	SCHOOL_ZONE	int64
FLAGS	SHLDR_RELATED	int64
FLAGS	SIGNALIZED_INT	int64
FLAGS	SINGLE_VEHICLE	int64
FLAGS	SNOWMOBILE	int64
FLAGS	SNOW_SLUSH_ROAD	int64
FLAGS	SPEEDING	int64
FLAGS	SPEEDING_RELATED	int64
FLAGS	SPEED_CHANGE_LANE	int64

Dataset	Column	Data Type
FLAGS	STATE_ROAD	int64
FLAGS	STOP_CONTROLLED_INT	int64
FLAGS	SUDDEN_DEER	int64
FLAGS	SUSPECTED_MINOR_INJURY	int64
FLAGS	SUSPECTED_SERIOUS_INJURY	int64
FLAGS	SV_RUN_OFF_RD	int64
FLAGS	TAILGATING	int64
FLAGS	TRAIN	int64
FLAGS	TRAIN_TROLLEY	int64
FLAGS	TROLLEY	int64
FLAGS	TURNPIKE	int64
FLAGS	UNBELTED	int64
FLAGS	UNDERAGE_DRNK_DRV	int64
FLAGS	UNLICENSED	int64
FLAGS	UNSIGNALIZED_INT	int64
FLAGS	URBAN	int64
FLAGS	VEHICLE_FAILURE	int64
FLAGS	VEHICLE_TOWED	int64
FLAGS	VULNERABLE_ROAD_USER	int64
FLAGS	VULNERABLE_ROAD_USER_FATAL	int64
FLAGS	WET_ROAD	int64
FLAGS	WORK_ZONE	int64
FLAGS	YOUNG_DRIVER	int64
PERSON	CRN	int64
PERSON	AGE	int64
PERSON	AIRBAG1	object
PERSON	AIRBAG2	float64
PERSON	AIRBAG3	float64
PERSON	AIRBAG4	float64
PERSON	AIRBAG_PADS	float64
PERSON	DVR_LIC_STATE	str
PERSON	DVR_PED_CONDITION	float64
PERSON	EJECTION_IND	float64
PERSON	EJECT_PATH_CD	float64
PERSON	EXTRIC_IND	float64
PERSON	INJ_SEVERITY	float64
PERSON	NON_MOTORIST	int64
PERSON	PERSON_NUM	int64

Dataset	Column	Data Type
PERSON	PERSON_TYPE	float64
PERSON	RESTRAINT_HELMET	float64
PERSON	SEAT_POSITION	float64
PERSON	SEX	str
PERSON	TRANSPORTED	str
PERSON	TRANSPORTED_BY	float64
PERSON	UNIT_NUM	int64
PERSON	VULNERABLE_ROAD_USER	int64
ROADWAY	CRN	int64
ROADWAY	ACCESS_CTRL	float64
ROADWAY	COUNTY	int64
ROADWAY	LANE_COUNT	float64
ROADWAY	OFFSET	float64
ROADWAY	RAMP	int64
ROADWAY	RDWY_ORIENT	str
ROADWAY	RDWY_SEQ_NUM	int64
ROADWAY	ROAD_OWNER	float64
ROADWAY	ROUTE	str
ROADWAY	SEGMENT	float64
ROADWAY	SPEED_LIMIT	float64
ROADWAY	STREET_NAME	str
TRAILVEH	CRN	int64
TRAILVEH	TRAILER_PARTIAL_VIN	object
TRAILVEH	TRL_SEQ_NUM	int64
TRAILVEH	TRL_VEH_REG_STATE	str
TRAILVEH	TRL_VEH_TAG_NUM	str
TRAILVEH	TRL_VEH_TAG_YR	object
TRAILVEH	TRL_VEH_TYPE_CD	float64
TRAILVEH	UNIT_NUM	int64
VEHICLE	CRN	int64
VEHICLE	AVOID_MAN_CD	float64
VEHICLE	BODY_TYPE	object
VEHICLE	COMM_VEH_IND	str
VEHICLE	DAMAGE_IND	float64
VEHICLE	DVR_PRES_IND	float64
VEHICLE	EMERG_VEH_USE_CD	float64
VEHICLE	GRADE	float64
VEHICLE	HAZMAT_IND	str

Dataset	Column	Data Type
VEHICLE	IMPACT_POINT	float64
VEHICLE	INS_IND	str
VEHICLE	MAKE_CD	str
VEHICLE	MODEL_YR	float64
VEHICLE	NM_AT_INTERSECTION	str
VEHICLE	NM_CROSSING_TCD	float64
VEHICLE	NM_DISTRACTION	float64
VEHICLE	NM_IN_CROSSWALK	float64
VEHICLE	NM_LIGHTING	str
VEHICLE	NM_POWERED	float64
VEHICLE	NM_REFLECT	str
VEHICLE	NON_MOTORIST	int64
VEHICLE	OWNER_DRIVER	float64
VEHICLE	PARTIAL_VIN	str
VEHICLE	PEOPLE_IN_UNIT	int64
VEHICLE	PRIN_IMP_PT	float64
VEHICLE	RDWY_ALIGNMENT	float64
VEHICLE	SPECIAL_USAGE	float64
VEHICLE	TOW_IND	str
VEHICLE	TRAVEL_DIRECTION	str
VEHICLE	TRAVEL_SPD	object
VEHICLE	TRL_VEH_CNT	float64
VEHICLE	UNDER_RIDE_IND	float64
VEHICLE	UNIT_NUM	int64
VEHICLE	UNIT_TYPE	int64
VEHICLE	VEH_COLOR_CD	float64
VEHICLE	VEH_MOVEMENT	float64
VEHICLE	VEH_POSITION	float64
VEHICLE	VEH_REG_STATE	str
VEHICLE	VEH_ROLE_CD	float64
VEHICLE	VEH_TYPE	float64
VEHICLE	VINA_BODY_TYPE_CD	str

AI Statement of Use

AI (*Chatbot App—AI Chatbot*) was used to troubleshoot errors, especially errors related to setting up the virtual environment. It was also used to look up code syntax and definitions of terms for better understanding. AI was not used for writing or coding. No blocks of text or code were copied from AI for use in this project.

Bibliography

Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowl. Inf. Syst.*, *51*(2), 339–367. https://doi.org/10.1007/s10115-016-0987-z

Anderson, D. (2017, February 18). Paul J. Miller, 21—PA |. *EndDD*. https://www.enddd.org/the-impact/paul-j-miller-21-pa/

Chatbot App—AI Chatbot. (n.d.). Retrieved February 1, 2026, from https://chat.chatbot.app/?model=gpt-5-mini

Chmielinski, K., Newman, S., Kranzinger, C. N., Hind, M., Vaughan, J. W., Mitchell, M., Stoyanovich, J., McMillan-Major, A., McReynolds, E., Esfahany, K., Gray, M. L., Hudson, M., & Chang, A. (2024, May). *The CLeAR Documentation Framework for AI Transparency*.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism – A new way of thinking about data science and data ethics that is informed by the ideas of intersectional feminism.* https://datafeminism.io/

Henderson, T. (2023, November 13). Less driving but more deaths: Spike in traffic fatalities puzzles lawmakers. *Pennsylvania Capital-Star*. https://penncapital-star.com/transportation-infrastructure/less-driving-but-more-deaths-spike-in-traffic-fatalities-puzzles-lawmakers/

Leppert, A. J. and R. (2024, November 12). Many Americans perceive a rise in dangerous driving; 78% see cellphone distraction as major problem. *Pew Research Center*.

https://www.pewresearch.org/short-reads/2024/11/12/many-americans-perceive-a-rise-in-dangerous-driving-78-see-cellphone-distraction-as-major-problem/

Magensky, M. (2021, August 16). *Statistics show pandemic causing dangerous driving habits*. WHP. https://local21news.com/news/local/statistics-show-pandemic-causing-dangerous-driving-habits

NHTSA. (n.d.). *Aggressive Driving and Other Laws* [Text]. Retrieved January 22, 2026, from https://www.nhtsa.gov/book/countermeasures-that-work/speeding-and-speed-management/countermeasures/unproven-further-evaluation/aggressive

NSC. (2020, July 21). *Motor Vehicle Fatality Rates Rose 23.5% in May, Despite Quarantines—National Safety Council*. https://www.nsc.org/newsroom/motor-vehicle-fatality-rates-rose-23-5-in-may-desp

PennDOT. (2025). *Data Dictionary and Field Constraints Tables*. https://gis.penndot.pa.gov/gishub/crashZip/Crash_Data_Dictionary_2025.pdf

PennDOT. (n.d.). *Distracted Driving*. Retrieved January 22, 2026, from https://www.pa.gov/agencies/penndot/traveling-in-pa/safety/traffic-safety-driver-topics/distracted-driving

PennDOT Crash Data Download. (n.d.). Pennsylvania Department of Transportation. Retrieved January 25, 2026, from https://experience.arcgis.com/experience/51809b06e7b140208a4ed6fbad964990

Pennsylvania Crash Information Tool. (n.d.). Retrieved January 25, 2026, from https://crashinfo.penndot.pa.gov/PCIT/welcome.html?TYPE=33554433&REALMOID=06-6e7f3fd5-ee48-4b22-8067-774762688650&GUID=&SMAUTHREASON=0&METHOD=GET&SMAGENTNAME=-

SM-E%2bZZmVPChvgeLWKfG1BgkUwXVkEogJGcRA5yj1Pa9OazD3%2fSbooshnw8PzBLhpiK&TARGET=-SM-https%3a%2f%2fcrashinfo%2epenndot%2epa%2egov%2f#

Sandino, D. (2025, April 3). *Shapiro Administration Urges Drivers to Put the Phone Down and Drive Safely*. https://www.pa.gov/agencies/insurance/newsroom/shapiro-administration-urges-drivers-to-put-the-phone-down-and-d

Tefft, B. C., Añorve, V., Woon, K., & Kelley-Baker, T. (2021). Travel in the United States Before and During the COVID-19 Pandemic. *AAA Foundation for Traffic Safety*. https://aaafoundation.org/travel-in-the-united-states-before-and-during-the-covid-19-pandemic/

Tefft, B. C., Villavicencio, L., Benson, A., Arnold, L., Woon, K., Añorve, V., & Horrey, W. J. (2022). Self-Reported Risky Driving in Relation to Changes in Amount of Driving During the COVID-19 Pandemic. *AAA Foundation for Traffic Safety*. https://aaafoundation.org/self-reported-risky-driving-in-relation-to-changes-in-amount-of-driving-during-the-covid-19-pandemic/

WANGPRATHAM, N. (2023, January 31). *Maximize Your Time Series Analysis with Python's Change Point Detection Tools*. Medium. https://medium.datadriveninvestor.com/maximize-your-time-series-analysis-with-pythons-change-point-detection-tools-39ce2bc63be

Wolman, J. (2020, July 30). *Pa. Roads were less congested during the pandemic, but were they safer?* PennLive. https://www.pennlive.com/news/2020/07/pa-roads-were-less-congested-during-the-pandemic-but-were-they-safer.html