Week 6 Progress Report

Logan Camacho, Sam Adebayo, and Jacob Gavin

***Abstract*—(Logan Camacho) This project intends to pinpoint road segments within the state of Pennsylvania that carry a significantly higher proportion of car accidents, with predictions based on past car crash data in conjunction with mapped out traffic volume data.**

I. Introduction (Logan Camacho)

ACROSS the 67 counties located within the Commonwealth of Pennsylvania span 120,000 miles of state and locally owned roads and highways. [1] Accidents are a daily occurrence in today's car-centric infrastructure of America, so tracking and reporting crash rates is a trivial matter. In order to maximize civilian safety and protection statewide, a more in-depth analysis is necessary when considering the daily traffic volume. We are proposing that to spread awareness on road safety and accident prevention, we will be using the Pennsylvania Department of Transportation (PennDOT) car crash data from 2024 in conjunction with PennDOT traffic and mapping data made with ArcGIS to find which areas have the worst crash rate in comparison to its daily traffic volume, what justifies the severity rate of the crashes in these areas, as well as using other factors such as speed limits and seasonal changes can reduce/increase the rate of accidents in these high-volume areas.

September 15, 2025

II. Purpose of the project (Jacob Gavin)

The purpose of this project is to perform a comprehensive analysis of traffic accident data in the 67 counties of Pennsylvania to uncover patterns and underlying trends that contribute to road danger. Although traffic accidents are recorded and summarized by the state, these reports often lack the context of traffic volume, seasonal factors, and environmental factors that can increase risk. By integrating PennDOT crash records with traffic volume data and geospatial tools such as ArcGIS, this project aims to identify road segments and intersections that are disproportionately prone to accidents.

Beyond simple identification of high-risk areas, this project seeks to examine the contributing variables that define these zones, such as speed limits, lighting conditions, weather and seasonal fluctuations, and temporal patterns such as time of day or weekday vs. weekend. The goal is not only to highlight where crashes are most frequent, but also to determine why these areas are more hazardous, and what features they share.

Ultimately, the purpose of the project is to generate actionable insights that can inform decision-making by transportation planners, engineers, policymakers, and even the drivers themselves. By translating raw data into meaningful patterns, the project aims to provide a framework for interventions that improve safety, inform infrastructure investment, and reduce injuries and fatalities on Pennsylvania roads.

III. Beneficiaries of the project (Jacob Gavin)

This project is intended to benefit drivers in Pennsylvania. Although some of the insights gained from this project may not by themselves be usable by PA drivers, the insights may help to inform state and local policymakers to make decisions that are in the best interest of PA drivers. Ensuring the safety of drivers is a non-partisan issue that is unlikely to be subject to bias by decision makers. For example, suppose it is discovered that there are twice the number of accidents on a particular intersection in the winter months than there are in the summer months. In that case, it may suggest that additional investment needs to be made into winterizing the roads, with perhaps better rock salt coverage, better lighting, or something else.

The goal of this project will be to uncover the underlying trends in accident data, to see if there are any actions that can be taken to the benefit of Pennsylvania drivers, to ensure their safety on roads throughout the state and throughout the seasons. In addition, emergency responders and public health officials could find value in the project's results. By anticipating accident hotspots and understanding the underlying causes, emergency services can optimize their preparation, potentially reducing response times and improving outcomes for accident victims.

The broader community, including businesses and residents of areas where high accident rates are, will benefit from safer roads and fewer accidents. Reduced crash rates can lead to lower insurance premiums and decreased traffic congestion. Overall, the approach of this project ensures that its benefits extend well beyond individual drivers, supporting the safety of anyone who lives or works in Pennsylvania

IV. Anticipated Technical Challenges (Sam Adebayo)

Even though we will be working with a smaller data set, which will be either all of PA or certain sections of PA, we will still face challenges regarding the sizes of our datasets; such as how we would map things like road size or traffic congestion, the different variables used for different accident and vehicle types, sorting out location data, etc. Despite this, we can still examine prior papers to see how to move forward, which will be presented here.

Looking at previous studies, we found that, while complex, solving issues and creating models related to an issue as complex as traffic and accident data is not impossible. One such example comes from [2] at Iowa University, where they took data across the state of Iowa and did the following: "We impose a spatial grid S on the study area, where each grid *s(i)* represents a $d \times d$ square region. For example, if *d = 5km*, the entire state of Iowa can be partitioned into 128 × 64 grids". While this may be a bit more intensive, it is definitely something we can consider doing, given that our dataset includes coordinates for the crashes, which we could potentially map in a similar fashion. Deciding where to go with that data should also be easier, since plotting it should give us a better representation of clusters of data and where we should focus. I believe that looking for and finding more research papers related to traffic and accident reports could help us narrow down tactics we could use to interpret and move forward with the data.

A. Additional Traffic Insights

After taking the time to review other articles related to our project proposal, we have gained various insights that we can apply from now on with our data. Although our topic appears to be well-researched, readers will often find that there are multiple main issues related to traffic and accident data that have several solutions.

Returning to the paper cited in our proposal [2], it used k-means clustering on spatial data to map out the sections of the road in its given dataset, which in this case was Iowa. Their actual model, on the other hand, is a modified type of neural network known as Long Short-Term Memory (LSTM), which in their words "The ConvLSTM model is a variation of LSTM to handle spatio-temporal prediction problems ... Each input feature of a ConvLSTM network is a three-dimensional spatio-temporal tensor, where the first two dimensions are the spatial dimension. Comparing with the original LSTM model, the input-to-state and state-to-state transitions of the ConvLSTM cell involve convolutional operations that output 3-dimensional tensors". [2] To explain this, they would take the dimensions of one of their road segments, input them into their neural network, and have the network learn how to classify more dangerous road segments while " forgetting" any data that may be inconsequential to doing so. Again, this is a method that we could implement using ArcGIS data, but combining both that and our accident data may prove to be a challenge.

To build on this, [3] uses a Lebanese car accident database to map accidents to their locations in Lebanon. Then, to go further, they end up creating crash event maps of Beirut, and for their final solution, they assign hazard values to a road based on segment length, the number of crashes, and their data's time frame. Although this isn't too technical, we may end up creating a similar feature if we utilize road-segmented data in our final project.

Lastly, [4] uses a much different approach. Using variables from their crash dataset, they decided to remove null values and make their set binary. Afterward they mapped them out, then converted the coordinates to pixel conversions, populated them with their crash data, and then sent that final image (a 3x3 matrix) into their model. While we have a dataset to work with, implementing this may require a significant amount of time and resources, especially since some of our variables aren't easily converted into binary attributes. Overall, there is still more literature review to be done; however, based on what we've seen so far, our project does have room for nuance within this space.

The more prominent problem arises from our data and is immediately apparent upon a quick glance. There are a multitude of variables that we are dealing with, from accident type to the types of vehicle involved, the time of day and location, the number of people injured or involved in the accident, and so on, and so forth. We obviously do not plan to use the raw datasets in their entirety, but having to narrow them down for our specific needs will be a large task on its own. Besides this, we may have to edit our final data set because one variable was shown to have no significance, or maybe we missed a variable that could have drastically altered the data for the better. There is a possibility that some of the data we have could be very redundant, which could also hurt any model we decide to make. Overall, we just need to carefully look over our data sets and select which variables we believe will have the largest impact in our model and go from there.

V. Initial Solution framework and Implementation plan (Logan Camacho)

Traffic data analysis is an area that our team has not previously explored, so we aim to gain a deeper understanding of the crash data and familiarize ourselves with ArcGIS mapping software to extract traffic risks that contribute to increased crash frequency. Understanding how these crash hotspots occur is imperative to determining where they have the highest likelihood of appearing on the road.

Not only does this involve combing through the datasets for the extractable features and any formatting errors, but also reading the associated PennDOT documentation to further our contextual foundation on the subject matter. With the correlation between tables, table joins, and organizational column additions, this alludes to our team setting the definitive features and response variables for each stage of testing. Once the variables are segmented, we should be entering week 6 of this project as the beginning of model training and looping between variable-model combinations before increasing scale.

To prepare the model for statewide applications, it is in our best interest to begin training using simpler models with fewer iterations before increasing model complexity and region area. This is combined with the variable interpretation previously mentioned to maximize model performance to scale. This performance-at-scale optimization, combined with the preliminary evaluation of models, will mark a positive milestone for week nine. The more time allotted to handling visualization plans for the final presentation, the better off our team will be throughout this project.

VI. Data Overview (Jacob Gavin)

A. Files and Levels of Observation

a) `CRASH_2024.csv`: Crash-level records keyed by `CRN`. Contains time, location, environment, roadway context, and severity/count fields. Approx. 110,814 rows (47 MB).

b) `FLAGS_2024.csv`: Crash-level engineered binary flags keyed by `CRN`. Mirrors `CRASH` cardinality. Approx. 110,814 rows (56 MB).

c) `ROADWAY_2024.csv`: Crash–roadway context with one or more rows per `CRN` (sequence). Includes lane count, speed limit, access control, route/ segment. Approx. 174,250 rows (14 MB).

d) `VEHICLE_2024.csv`: Unit-level vehicle records keyed by `CRN` and `UNIT_NUM`. Vehicle type, movement, speed, roles, special usage. Approx. 197,507 rows (36 MB).

e) `PERSON_2024.csv`: Person-level records keyed by `CRN`, `UNIT_NUM`, `PERSON_NUM`. Demographics, role, restraint/helmet, injury severity. Approx. 244,174 rows (26 MB).

f) `COMMVEH_2024.csv`: Commercial vehicle details keyed by `CRN`, `UNIT_NUM`. GVWR, hazmat, carrier info. Approx. 8,717 rows (1.6 MB).

g) `CYCLE_2024.csv`: Motorcycle-related safety gear and attributes keyed by `CRN`, `UNIT_NUM`. Approx. 3,427 rows (308 KB).

h) `TRAILVEH_2024.csv`: Trailer details keyed by `CRN`, `UNIT_NUM` (with trailer sequence). Approx. 5,180 rows (312 KB).

B. Key Relationships

- Global crash identifier: `CRN` links all files.
- Vehicle joins: `CRN` + `UNIT_NUM`.
- Person joins: `CRN` + `UNIT_NUM` + `PERSON_NUM`.
- Roadway context: one or more records per `CRN` (sequence).

VII. Schema Summaries (Jacob Gavin)

A. `CRASH_2024.csv` (Crash-Level)

- Location/time: `DEC_LATITUDE`, `DEC_LONGITUDE`, `COUNTY`, `MUNICIPALITY`, `CRASH_MONTH`, `DAY_OF_WEEK`, `HOUR_OF_DAY`, `TIME_OF_DAY`.
- Environment: `WEATHER1`, `WEATHER2`, `ILLUMINATION`, `ROAD_CONDITION`, `RDWY_SURF_TYPE_CD`.
- Control/geometry: `TCD_TYPE`, `TCD_FUNC_CD`, `INTERSECT_TYPE`, `INTERSECTION_RELATED`, `RELATION_TO_ROAD`, `LOCATION_TYPE`.
- Work zone: `WORK_ZONE_IND`, `WORK_ZONE_TYPE`, `WORK_ZONE_LOC`, `WZ_*`.
- Counts/severity: vehicle/person counts, `MAX_SEVERITY_LEVEL`, injury/fatal counts.

B. `FLAGS_2024.csv` (Crash-Level Flags)

- Behaviors/context/crash types: e.g., `ALCOHOL_RELATED`, `DRUG_RELATED`, `DISTRACTED`, `SPEEDING_RELATED`, `INTERSECTION`, `LANE_DEPARTURE`, `REAR_END`, `LEFT_TURN`, `OVERTURNED`, `WORK_ZONE`.
- Outcome-adjacent flags: `FATAL`, `INJURY`, `FATAL_OR_SUSP_SERIOUS_INJ` (exclude as predictors when used as labels).

C. `ROADWAY_2024.csv` (Crash–Roadway Context)

- Design/operations: `LANE_COUNT`, `ACCESS_CTRL`, `SPEED_LIMIT`, `RDWY_ORIENT`, `RAMP`, `ROAD_OWNER`.
- Network IDs: `ROUTE`, `SEGMENT`; enables segment-level aggregation.
- Cardinality: multiple rows per `CRN` possible.

D. `VEHICLE_2024.csv` (Unit-Level)

nosep

- Attributes: `UNIT_TYPE`, `VEH_TYPE`, `BODY_TYPE`, `VEH_COLOR_CD`, `MODEL_YR`, `COMM_VEH_IND`.
- Dynamics/context: `VEH_MOVEMENT`, `TRAVEL_SPD`, `IMPACT_POINT`, `PRIN_IMP_PT`, `TOW_IND`, `SPECIAL_USAGE`.
- Non-motorist interaction indicators included.

E. `PERSON_2024.csv` (Person-Level)

- Demographics/roles: `AGE`, `SEX`, `PERSON_TYPE`, `SEAT_POSITION`, `NON_MOTORIST`.
- Safety/condition: `RESTRAINT_HELMET`, `AIRBAG1–AIRBAG4`, `EJECTION_IND`, `INJ_SEVERITY`, `TRANSPORTED`.

F. `COMMVEH_2024.csv`, `CYCLE_2024.csv`, `TRAILVEH_2024.csv`

- Commercial/heavy truck: GVWR, hazmat indicators, carrier identifiers.
- Motorcycle: helmet and protective gear, engine size.
- Trailer: presence and type; multiple trailers per unit possible.

VIII. Cleaning Plan (Jacob Gavin)

A. Global Ingestion and Normalization

1) Read CSVs with consistent encoding; trim whitespace; preserve provided column names.
2) Enforce types:
   - Identifiers: `CRN` as string-like key; `UNIT_NUM`, `PERSON_NUM` as integers.
   - Categorical: codes (e.g., `TCD_TYPE`, `ILLUMINATION`, `UNIT_TYPE`) as categorical strings.
   - Numeric: counts, speeds, limits, lane counts to numeric with coercion.
   - Binary flags: coerce to 0/1 integers.
3) Standardize missing values: convert blanks and placeholders (e.g., `NA`, `NULL`) to NA.
4) Deduplicate: ensure `CRN` uniqueness at crash level; drop exact duplicates elsewhere.
5) Geospatial checks: validate latitude/longitude ranges for Pennsylvania; flag out-of-bounds.
6) Temporal checks: verify `CRASH_YEAR`=2024; standardize month/day/hour fields.
7) Category harmonization: document codebooks for categorical values.

References

[1] Pennsylvania Department of Transportation, “2024 pennsylvania crash facts & statistics,” Bureau of Operations – Crash Information Systems and Analysis Unit, Harrisburg, PA, Tech. Rep., 2025, accessed: 2025-09-11. [Online]. Available: https://www.pa.gov/content/dam/copapwp-pagov/en/penndot/documents/travelinpa/safety/documents/2024_cfb_linked.pdf

[2] Z. Yuan, X. Zhou, and T. Yang, “Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’18. New York, NY, USA: Association for Computing Machinery, 2018, p. 984–992. [Online]. Available: https://doi.org/10.1145/3219819.3219922

[3] A. J. Ghandour, H. Hammoud, and L. Telesca, “Transportation hazard spatial analysis using crowd-sourced social network data,” *Physica A: Statistical Mechanics and its Applications*, 2019, accessed: Oct. 2, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437119300251

[4] A. Rahim and H. M. Hassan, “A deep learning based traffic crash severity prediction framework,” *Accident Analysis & Prevention*, 2021, accessed: Oct. 2, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457521001214