Mid-Point Check In- Jared Zirkes and Nadav Gerner

Part I

As we approach the mid-point check-in for our final project, almost everything from our initial project proposal remains on track. At a high level, the project goal is to explore the presence of Car Crashes in the District of Columbia, while answering a few specific questions about the overall 'landscape' of crashes. We are accomplishing this task with a main dataset of all crashes, as compiled by the District, which contains both general data about each crash, like parties injured, and circumstances surrounding the crash, but also geographic Latitude/Longitude data for each crash present, allowing us to explore the crashes geographically. The only change since our initial proposal is the inclusion of proxy data to normalize the 'danger' presence for different geographic segments. This is done via data on the number of owned cars per District Ward as well as geographic Capital Bikeshare data, which will allow us to normalize the crash volumes with the total volume of traffic. Now, we will be able to show which areas are hotspots not only for the number of crashes, but which are really the most dangerous, giving a sense of 'danger'.

Part II

As we have begun working through all our available data, we have been able to begin formalizing and crafting the final narrative. The narrative we want to follow is one following this linear structure:

1. Setting up the stage: high-level crash trends in the city.
2. Crash hotspot: visualization of what leads to a crash hotspot.
3. Fatal crashes overview: high-level fatal crash trends in the city.
4. Fatal hotspots: visualizations exploring fatalities and differences from other non-fatal crash hotspots.
5. Temporal nature of accidents in the District: mostly around fatal accidents here, where we will explore using time series and other visualizations capturing accidents and their changes in trends over time.

As for the expected format of our deliverables, we really want it to follow our narrative and to emphasize a linear storytelling method. This will allow us to explore the major questions in order, 'zooming' in and out of the big picture to focus on the pieces we feel most compelling. This, in turn, allows us to control the story for the data consumer.

As we are still working on creating these visualizations, these are subject to change in the details. With that, we have a strong understanding of what we want to present and visualize to help our storytelling.

First, we want to start with a visualization that will help us answer: "What are the hotspots for total crashes in DC?" and also "Does normalizing crashes by traffic volume (thus localizing dangerous spots) change those hotspots?" This will give an overview of the traffic and crashes in DC, giving the consumer important information and also necessary context for further visualizations. We want to incorporate a choropleth map of DC, with a corresponding heatmap showing "danger" areas layered on top. This visualization will be interactive. The heatmap would be changed based on different types of "danger", and the corresponding areas will be able to be changed accordingly. The filtering we wanted to start with is a selection between total crashes and normalized crashes to traffic volume.

As we wanted to “zoom” in and out metaphorically to different focus areas (so not on the map, but rather areas we saw on the map and dug into). We wanted to zoom in on the most dangerous area we will find by the total number of accidents to answer: “Why is this area a hotspot for crashes?”. Zooming into the specific area, we wanted to create a second “view”. This visualization will be a static one, showing different visualizations within. We planned on a subplot containing a histogram/density plot of the volume of cars and/or accidents per time frame (likely year). In addition to a subplot giving insight into many types of accidents that occurred in this area, and other metadata of the crashes, such as what time of day these accidents happen, whether speeding was a factor, and whether bikers were involved. This data will allow us to further probe the question of ‘why’ this area might be a dangerous one for accidents.

Now, we wanted to zoom out and have another choropleth visualization showing the fatal crashes in DC. We expect some overlap, but the main area for fatal crashes should be different from high-crash areas often involve minor collisions, yet fatal crashes are more likely to occur on roads with higher speed limits or less congestion. Using this visualization, we could answer the question ‘where are the local Fatal hotspots?’. Here, we will want to incorporate an innovative approach; a “regular” map visualization with a heatmap will be repetitive. We are brainstorming ideas still, but we do want it to be linked to visualizations as well. Perhaps

“What makes a fatal crash hotspot? Why are there so many fatal crashes here?”. These are the questions we will want to answer in the next view, where we “zoom” in and explore these hotspots of fatal crashes. We will try to answer the questions with different plots between fatalities and other features, such as car and bike traffic, as well as contributing/present factors such as speeding and bikers. Here we will explore weather correlation visually, as days with extreme weather conditions might yield a higher number of fatalities.

The last view will be used to explore the temporal nature of accidents in the District. In this view, we will explore questions such as ‘Have the fatality of accidents changed over time?’. We can discuss this question by looking at normalized fatalities by year as well as the annual share of crashes that are fatal. From a plotting standpoint, we plan to utilize both line charts that easily visually capture the temporal nature of the data, as well as a map with encodings that capture changes over time, such as color coding the YoY differences in fatalities and/or fatality share. This will be a linked view visualization, where you will be able to filter the fatalities based on the timeline in the line plots.

In terms of design choices, For static visualizations, we created a “darker” background theme with colors that reflect urgency, danger, and threat. Colors like red and orange (shades that are easily distinguishable) will do that best. We will use lighter colors like light blue to represent safety and calm. Together with the darker background, they will contrast well, which will help us convey messages faster.

Thinking about how we will need to transform/convert our datasets into clean representations for visualizing, the process is thankfully quite straightforward. There will be three major steps we need to take. First, given the polygon data for each ward, which we plan to use as the baseline geographic separation, we will need to classify each event (Crash, and Bike ride) as belonging to an individual ward. This is as simple as writing a function to check the geography based on the latitude and longitude of each event and assigning it to the bounds it

falls in. Then, the second step is aggregating our car and bike data to determine the overall usage or presence by time period (in our case years). Finally, to condense everything into one final dataset, we will need to join all of the data based on the occurrence time, whether it be day or year, thus, assigning each crash a ward, a normalization value (from bike data and/or car ownership data), and weather data from the date of the crash.