For this project, we used a dataset provided by the course. We used the example dataset. It consists of 38 columns and 194,673 collisions. Each training example is a vehicle accident corresponding to one accident, which may or may not involve numerous people, including drivers, passengers, and pedestrians, and/or vehicles, including cars and bicycles.

For the model we use the following attributes: weather, junction, pedestrians, collision type and its description, light, road condition, and vehicle count.

For weather, there are 10 unique labels: 'Raining' 'Overcast' 'Clear' 'Unknown' 'Other' 'Snowing' 'Fog/Smog/Smoke' 'Sleet/Hail/Freezing Rain' 'Severe Crosswind' 'Blowing Sand/Dirt'. For junction, there are 6 unique labels: 'Mid-Block (not related to intersection)' 'At Intersection (intersection related)' 'Driveway Junction' 'At Intersection (but not related to intersection)' 'Mid-Block (but intersection related)' 'Ramp Junction'. For pedestrians, there are 7 unique labels: 0, 1, 2, 3, 4, 5, 6. For collision type, there are 10 unique labels: 'Sideswipe' 'Parked Car' 'Angles' 'Rear Ended' 'Left Turn' 'Other' 'Pedestrian' 'Right Turn' 'Cycles' 'Head On'. For light, there are 8 unique labels: 'Dark - Street Lights On' 'Daylight' 'Dark - No Street Lights' 'Unknown' 'Dusk' 'Dark - Street Lights Off' 'Other' 'Dawn'. For road condition, there are 9 unique labels: 'Wet' 'Dry' 'Unknown' 'Snow/Slush' 'Ice' 'Sand/Mud/Dirt' 'Standing Water' 'Other' 'Oil'. Finally, for vehicle count, there are 12 unique labels: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11. For our dependent variable, severity code, there are 2 unique labels: 1 (when only property was damage), and 2 (when the collision resulted in a non-serious injury)

These attributes were chosen because they displayed obvious trends related to our dependent variable, severity code. For example, let's look at the junction type vs. severity code in the bar plot to the right. You tell, just by looking at this graph that an accident is more likely to cause an
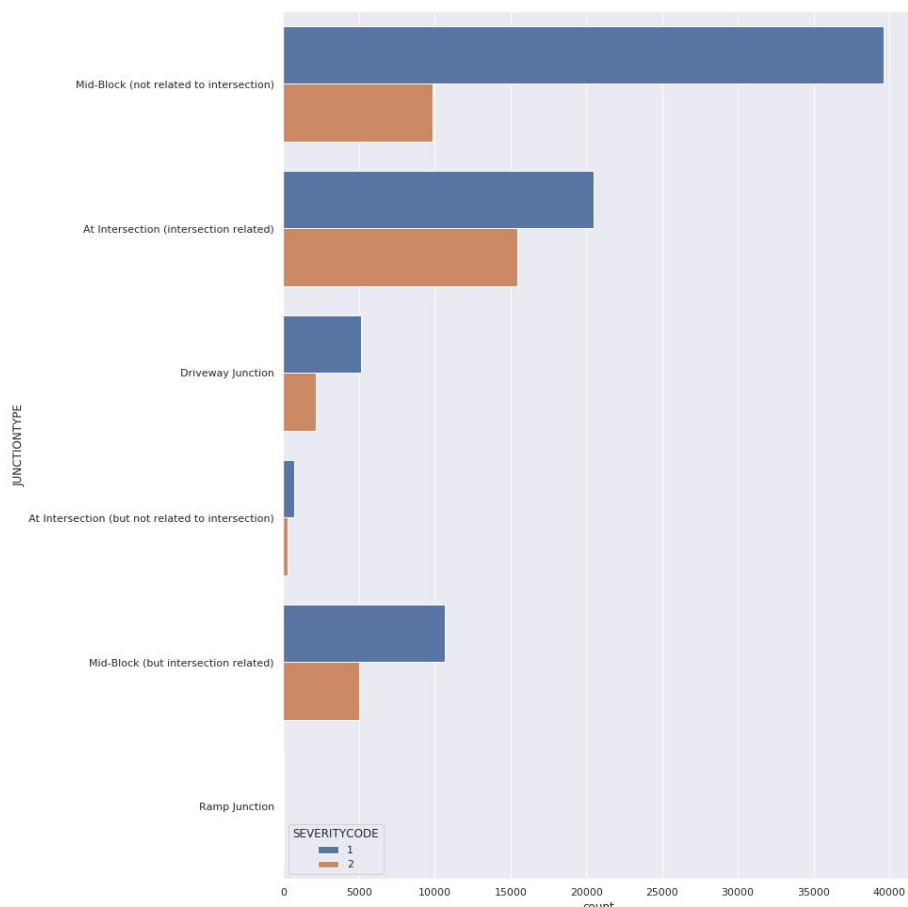
injury if it occurs at an intersection and the cause can be attributed to the intersection -- see At Intersection (intersection related). This is further evidenced by Mid-Block (but intersection related). We can quickly come to the belief that an accident that is related to an intersection in some way will result in an injury more often than an accident that is unrelated to an intersection.

Let's take a look at another example. See the figure below. This is a figure comparing the type of collision and its severity code.
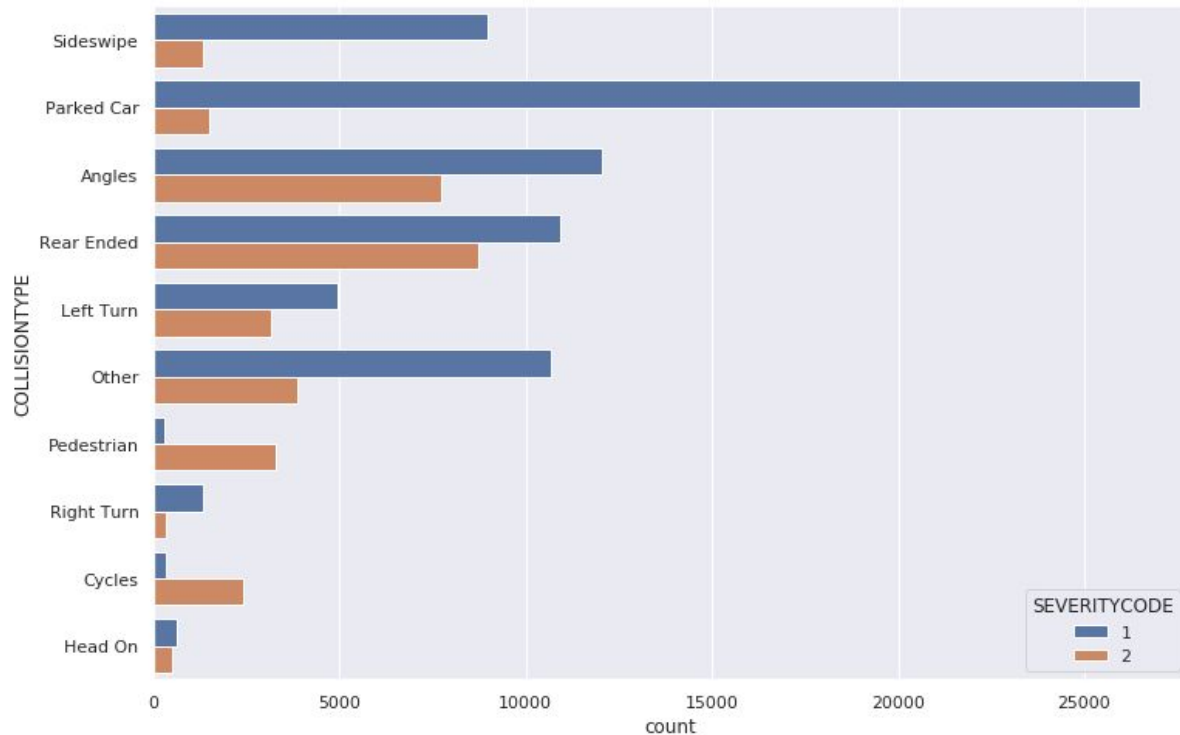
By looking at the image, one can reasonably conclude that an accident is more likely to cause an injury (severity code of 2) if the collision involves angles, being rear ended, a left turn, a pedestrian, a cyclist, or is head on. In contrast, you can tell it's less likely to cause an injury if the collision is a sideswipe, involves a parked car, or a right turn.