Slide 10

After briefly examining the datasets in excel to determine how many rows and columns, and what each column held, we loaded the two raw datasets into a postgreSQL database and used pgAdmin to further explore, clean, and ultimately combine the datasets for the machine learning model.  We first explored the raw\_vehicle\_traffic table, determining that it had an entry for every hour of every day from 2012 to 2018, with a Temperature in kelvin, rain and snow in mm per hour, a percent of cloud cover, two columns describing the weather conditions in text, and a total traffic volume for the hour, and a column listing what holiday, if any, a date was.  Upon closer inspection, we realized that several hours had duplicate entries, with duplicate weather and traffic values and slightly differing descriptions.  We dropped these descriptive columns and selected a distinct entry for each datetime and weather values.  There were still a few datetimes with multiple entries and slightly different weather measurements.  We averaged the weather measurements in those cases to get a single unique entry per datetime.

Slide 11

Next we found that for holidays, only 1 hour had the holiday labelled per day, instead of every datetime.  To correct this, we dropped the holiday column from the vehicle traffic table and created a new holiday table with each holiday and the date, which could later be joined back to the rest of the data to apply the holiday label correctly to all corresponding dates.

Slide 12

In the raw bike-pedestrian table, there were several columns with metadata describing sites and technology used for data collection.  There sites in multiple counties throughout Minnesota; we only needed data for Ramsey county where the vehicle traffic was measured, and did not need the metadata nor the weather data, as we had weather data already in the other table.  The data was recorded daily, not per hour as in the other table.  We checked to make to see if each date was unique, finding that there were multiple entries for Ramsey county--an entry for pedestrian and one for bike, and there were two different sites measuring.  Since we wanted to study all non-vehicle traffic, we summed the traffic from both modes to get a single total for each date.

Slide 13

Because the vehicle traffic table had entries for each hour while the pedestrian table only had entries for day, we had to create date column and group by date in the vehicle table, summing the rain, snow, and traffic for each day and getting an average temperature and cloud cover.  We also converted the Kelvin temperature to Farehnheit, then joined the vehicle, pedestrian, and holiday tables to create a single dataset to use in the machine learning model.  Since we were already adding a date column derived from the datetime at this step, we also added a day of week and month of year column extracted from the date as a preliminary part of feature engineering.

Slide 14

The vehicle traffic dataset contained data on traffic and weather for every hour of the day. However, because our pedestrian dataset contained only information on the amount of bikers on a given day, we had to collapse our vehicle dataset into daily categories instead of hourly categories in order to join the two datasets together. This caused us to lose insights into more nuanced vehicle traffic data, such as how the time of day affected vehicle traffic.

Also, the vehicle traffic only collected data from one road in Ramsey County, whereas the pedestrian traffic data was collected from two different locations in Ramsey County. Although the data points were collected from two different locations, we believe it still conveys a useful depiction  of pedestrian and vehicle traffic in the county.