Slide 10

After briefly examining the datasets in excel to determine how many rows and columns, and what each column held, and seeing that temp was in Kelvin, rain and snow in mm per hour, and clouds a, 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, and found that while there should be 1 entry per date per hour, some date\_times had up to six entris, 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 hour in the same day, as you can see on the left where October 8 at Midnight is labelled Columbus day, while 1,2,3, etc are labelled ‘None’.  To correct this, we dropped the holiday column from the vehicle traffic table and created a new holiday table shown on the right 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, unlike the vehicle set, there was only data on traffic per day, not per hour. There were several columns with metadata describing sites and technology used for data collection as well as some weather columns.  We did not need this metadata nor the weather readings since we had more complete data from other table. In this dataset, found entries for multiple counties in Minnesota. When filtering to just Ramsey county, to match the location of the data in the vehicle dataset, we found Ramsey county had two sites, one north/south, one east/west, and separate entries for each site for pedestrian and bike traffic, but unlike the other dataset, there were no duplicate entries per date. Since we wanted a total for all non-motor-vehicle traffic in the area, we aggregated the totals for each site and mode of transportation for each day.

Slide 13

Because the vehicle traffic table had hourly entries while the pedestrian table only had daily entries, 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.