Bike Speed and Weather

Mark Holt

24 Nov 2014

Bike Speed and Weather

Project for Data Science with R: Data Analysis

Acknowledgements:

- **Citibike operated by NYC Bike Share**

  provided all primary data

- **MapQuest Open Platform Web Services**

  provided all distance and elevation data

- **Weather Underground**

  provided all weather information

Aims

- To collect and integrate one years data from disparate data sources
- To estimate the bike speeds of the fastest riders
- To relate temperature and elevation with bike speed
- To predict journey time given weather conditions

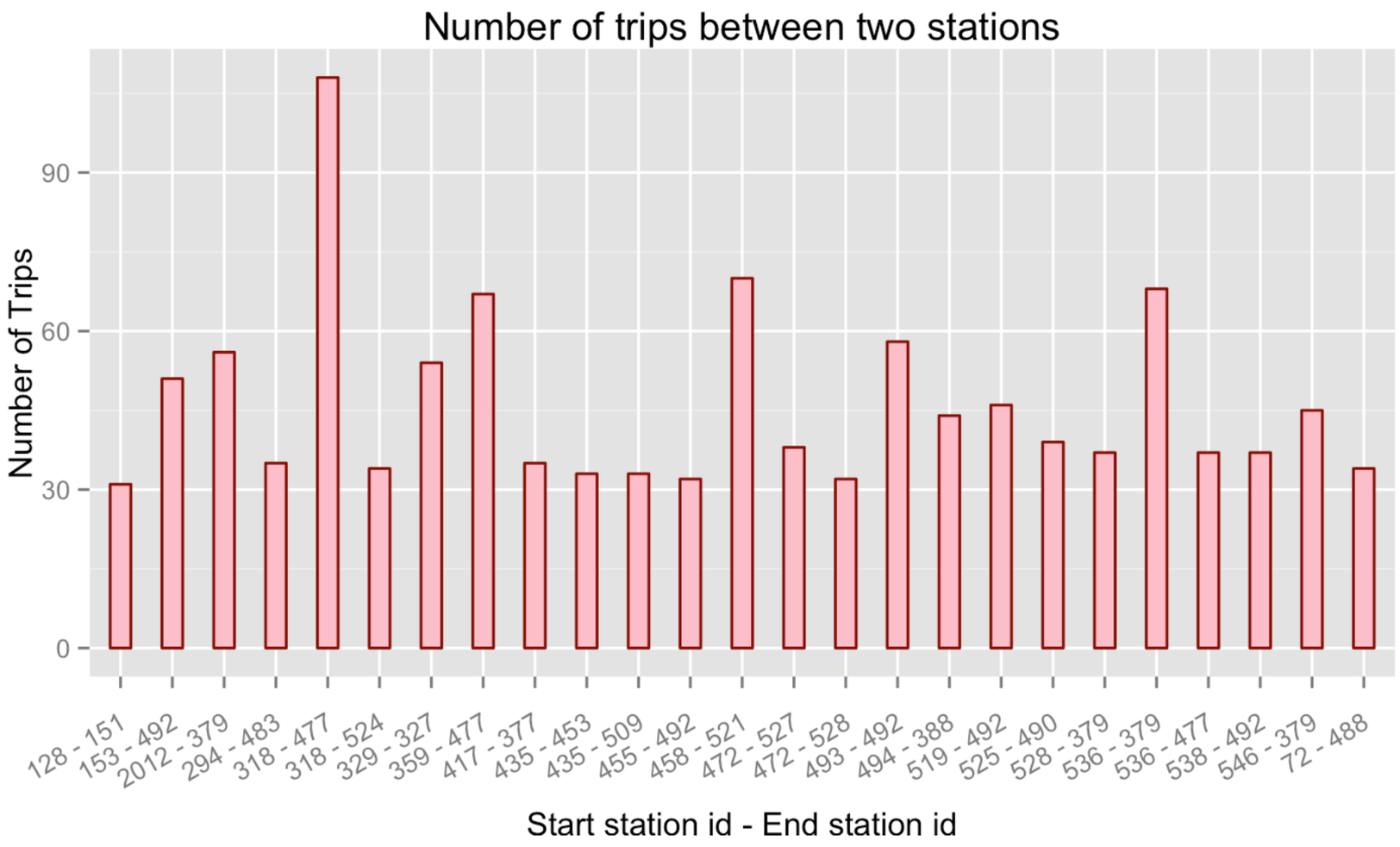
Citibike Data

- Substantial data available
- start station: id, lat, long, address
- end station: id, lat, long, address
- trip: start time, end time, trip duration, bike id
- user: customer or subscriber, birth year, gender

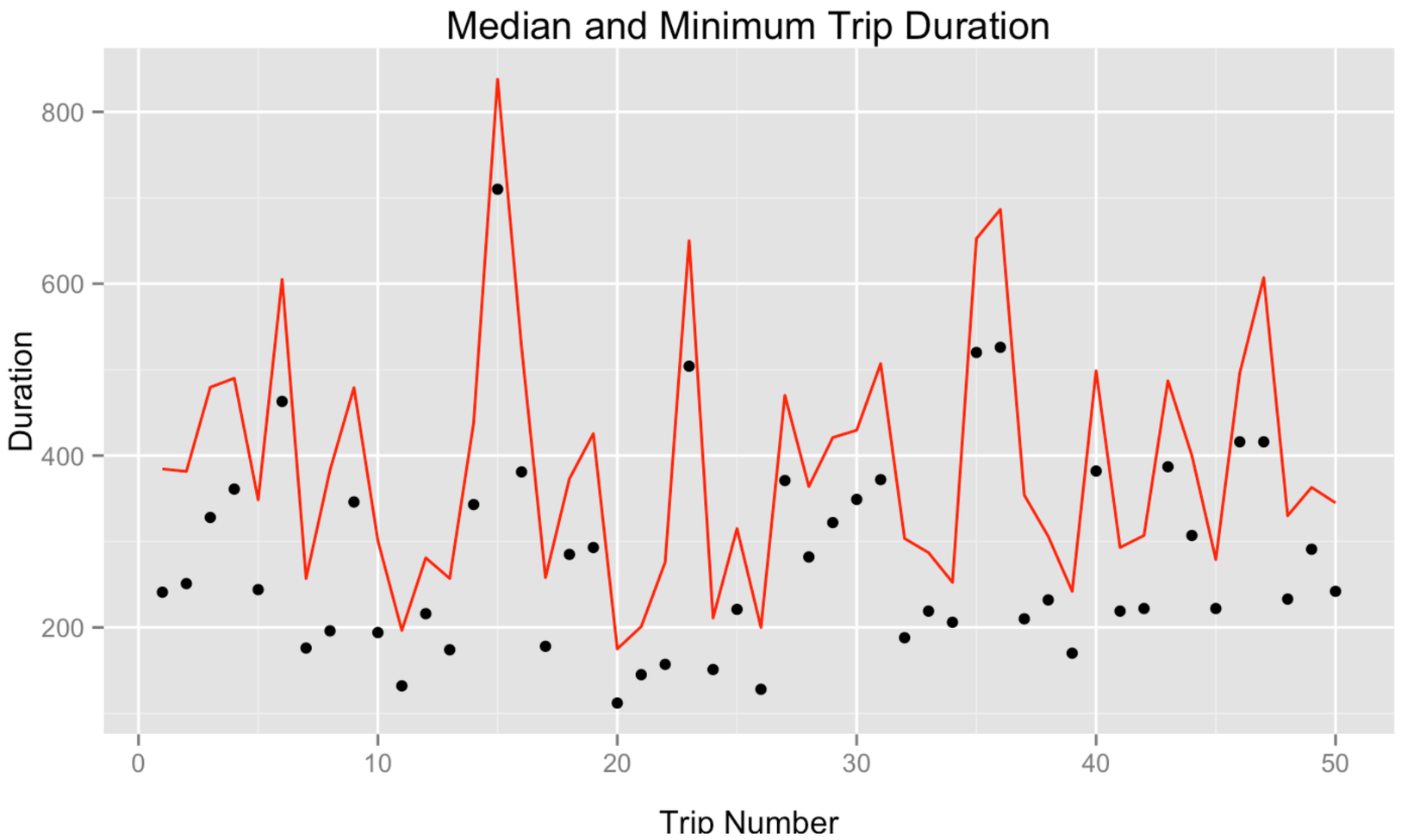
Data Pre-processing 1

- Filter the citibike data:
- Men, Subscribers, Mon - Fri, 9.00 - 5.00pm, Workdays only (excluded holidays)
- Find the fatest riders for all the station pairs

For Each Months Data



Finding the Fatest Cyclist



Data Pre-processing 2

- provides a "distance" from starting lat/long to ending lat/long
- From this data derive bike speed (in mph)
- Second call provides elevation bewtween start and end stations
- returns historical weather data close to the starting time of the trip
- Obtain observations for temp, windspeed, windchill, precipitation, humidity

Do the derived bike speeds seem credible?

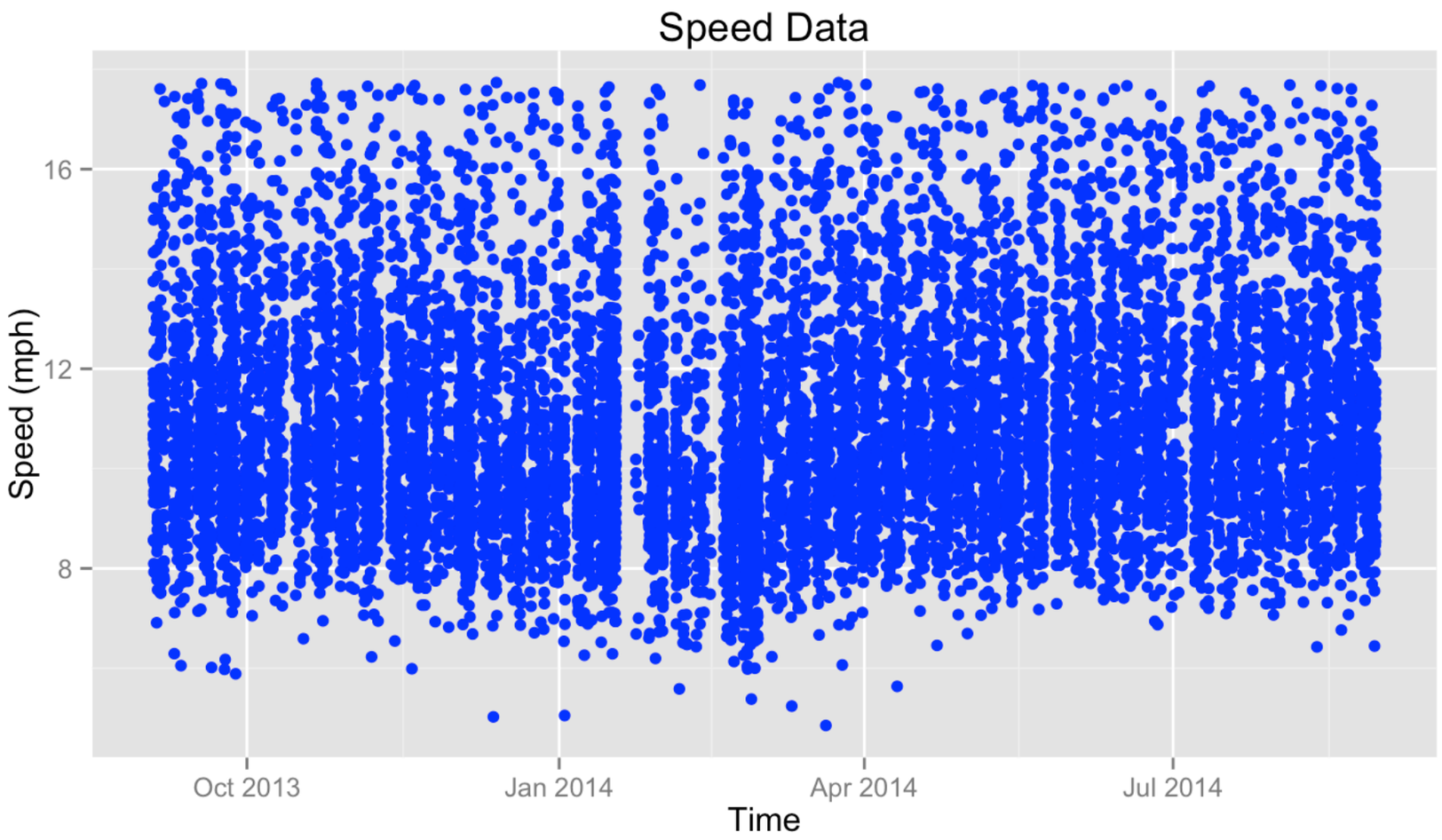
- Pablo Jensen at the École Normale Supérieure de Lyon, found:

  Over an average trip, cyclists travel 1.55 miles in 14.7 minutes

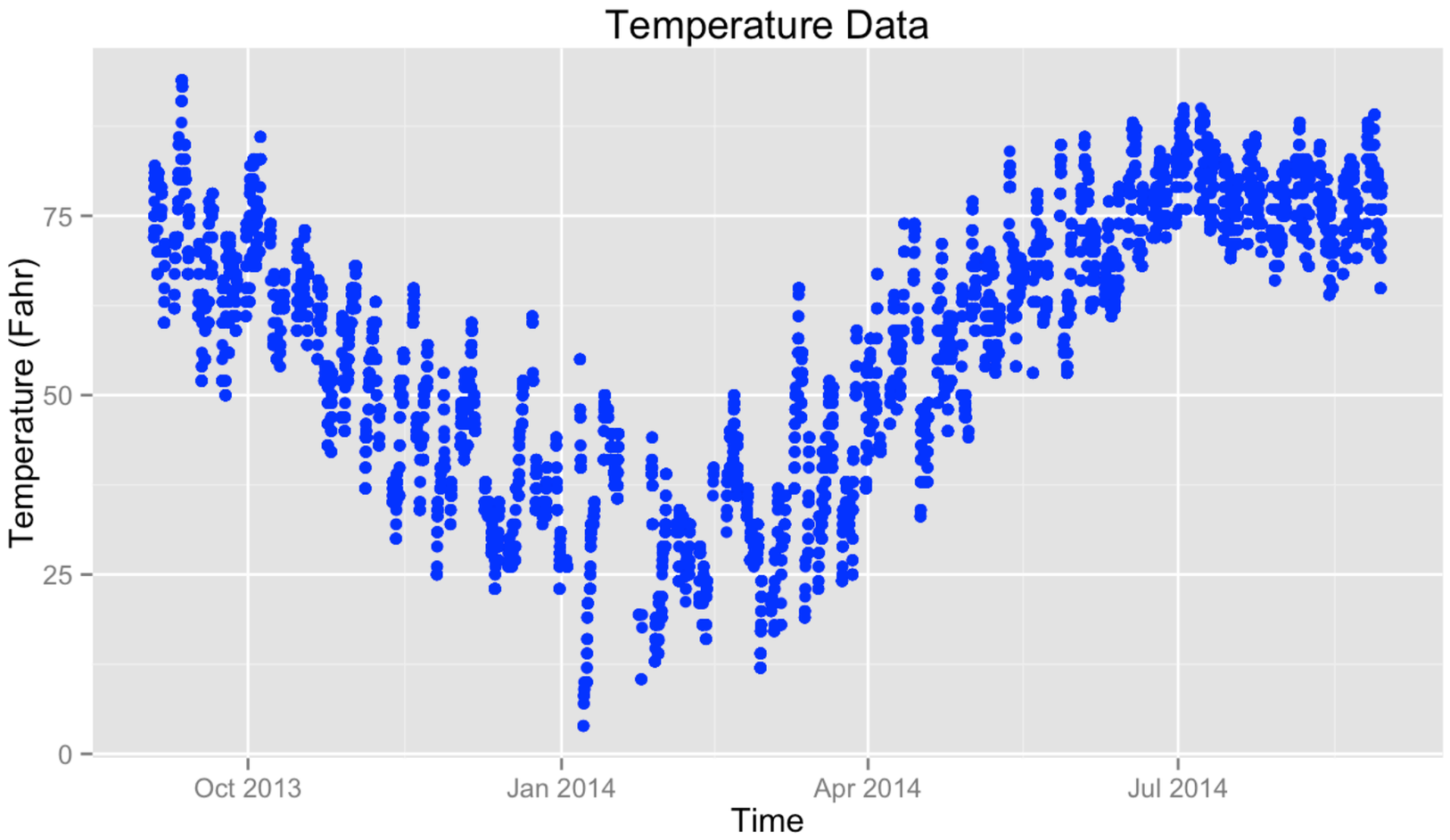
  Average speed ~ 6.2 mph.

  Rush hour average speed ~ 9.3 mph
- Citibike data speed estimates ~ 10.6 mph
- Perhaps the MapQuest route over-estimate the distance.

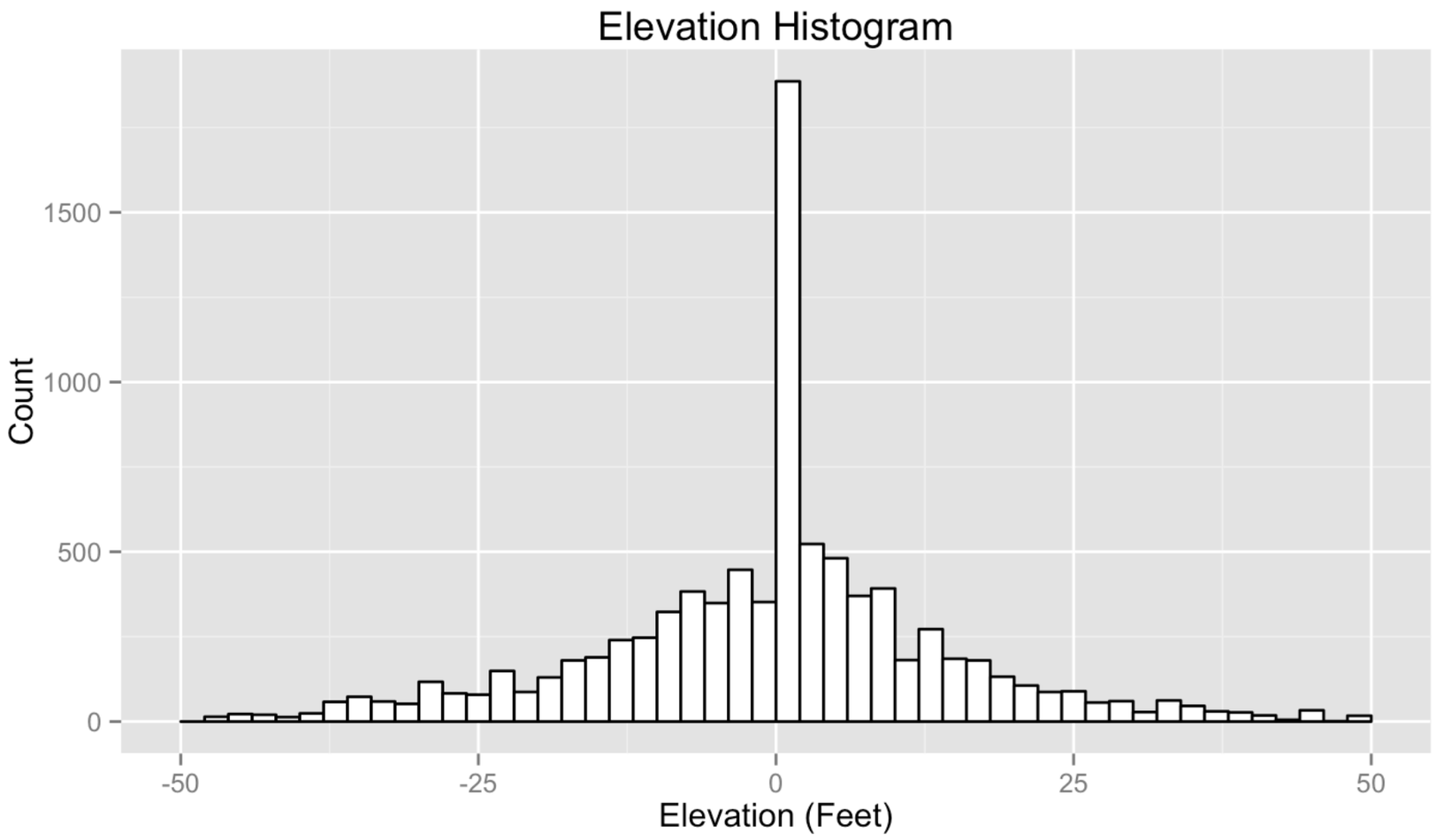
Data Visualization 1 - Speed



Data Visualization 2 - Temp



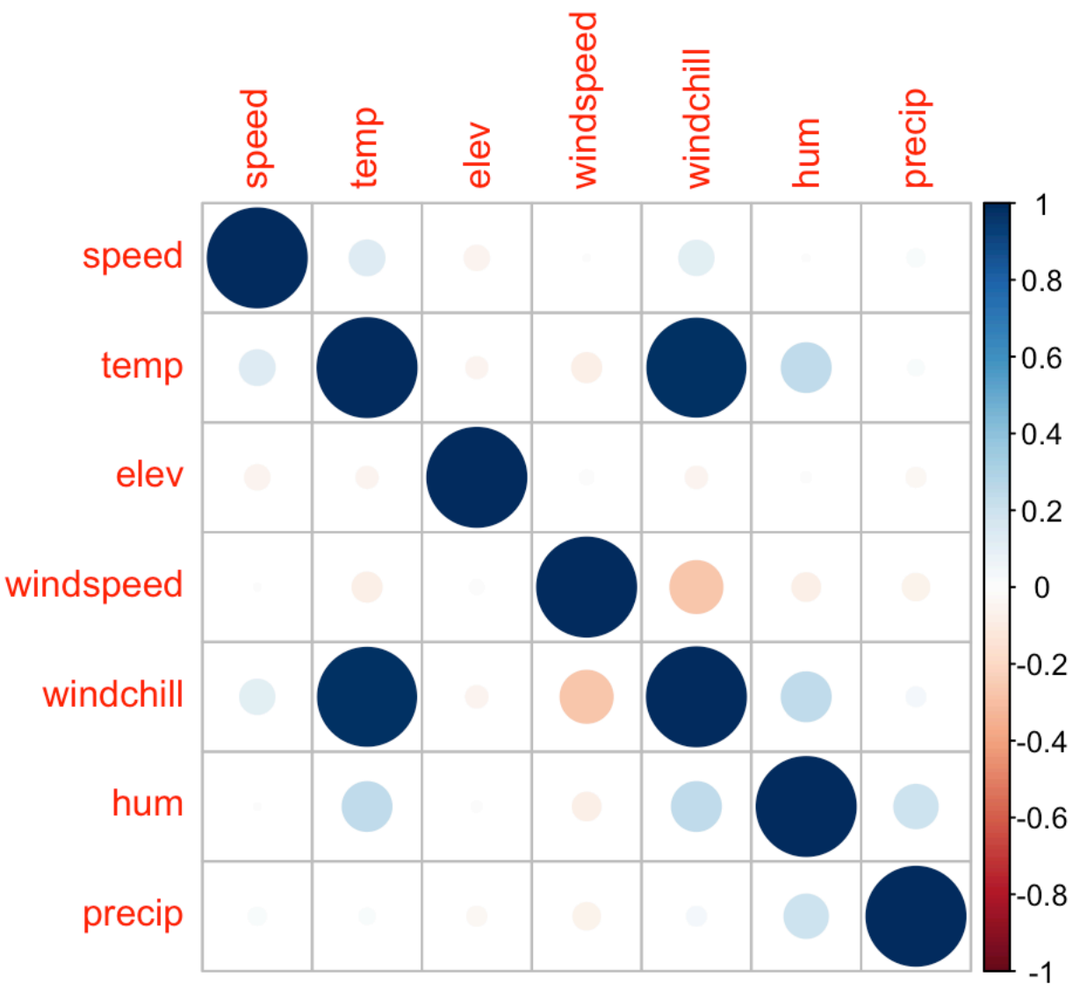
Data Visualization 3 - Elevation



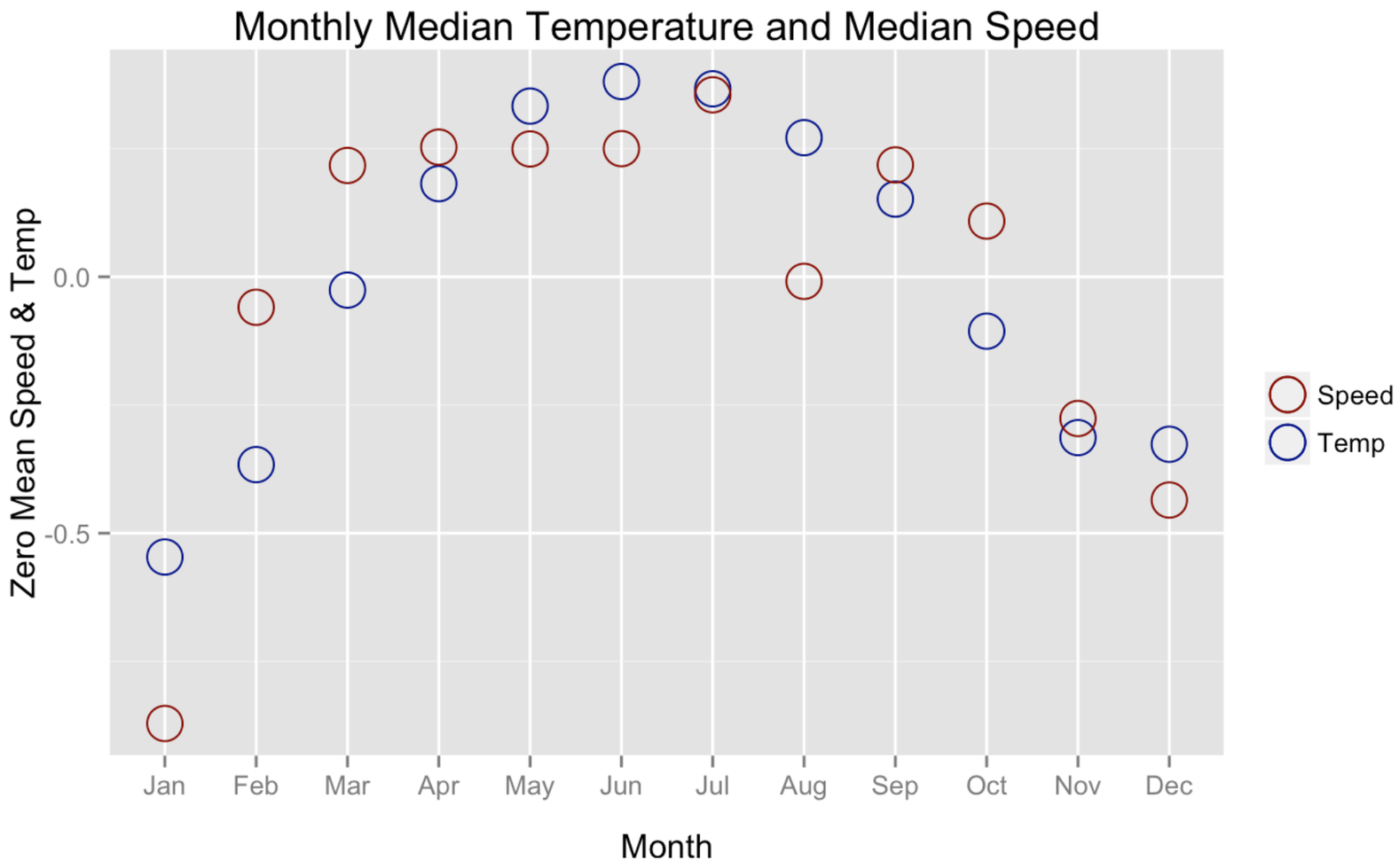
Other Weather Variables?

- Hoped to use precipitation (rain), but the data reported from Wunderground was sparse.
- Public blogs suggest they have a bug in their historical data reporting for rainfall!
- Windspeed: Data sparsely reported. Mostly reported as 0 mph.
- Humidity: Correlated more with temperature
- Windchill: Data sparsely reported. Correlated with temperature

Pairwise correlations



Speed and Temperature



Selecting the Model

- Randomly split data into training and test set (50:50)
- BoxCox plot: transform speed into log(speed)
- Normalized the data (zero mean and unit variance)
- Tested 5 models relating speed to temperature and elevation
- "Best" model: Log(Speed) ~ temp + temp-squared + elev
- "Best" meaning lowest MSE on the unseen test data, and lowest AIC
- MSE Test = 0.99

Using the Model

- Plan your trip
- Example: E34 & Vanderbilt to 11th & W27
- Input the station id's: 318 & 458
- Use MapQuest Api to get distance and elevation: 1.942 miles, 0 feet
- Use Wunderground Api to get low and high forecast for the day: 23 F & 36 F for Tuesday 18th Nov 2014
- Use the model to predict the speeds: 10.1 mph & 10.4 mph
- Estimate trip times: 11 mins & 11.5 mins

Conclusions

- Speed and temperature are correlated, but does the temperature in a different speed?
- Temperature is known to be very important to physiology.
- Incline and speed are related, but inclines in NYC are minimal
- The route of each cyclist is not actually known but speed can be approximated
- It would be interesting to know the true speeds of the fastest cyclists
- Data from disparate sources can be utilized in an integrated and therefore meaningful manner