Bike Sharing Dataset

Stephen Leonard

Initial Inspection

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349

Date

Weather Properties

Rider count

0 nulls in data, very clean!

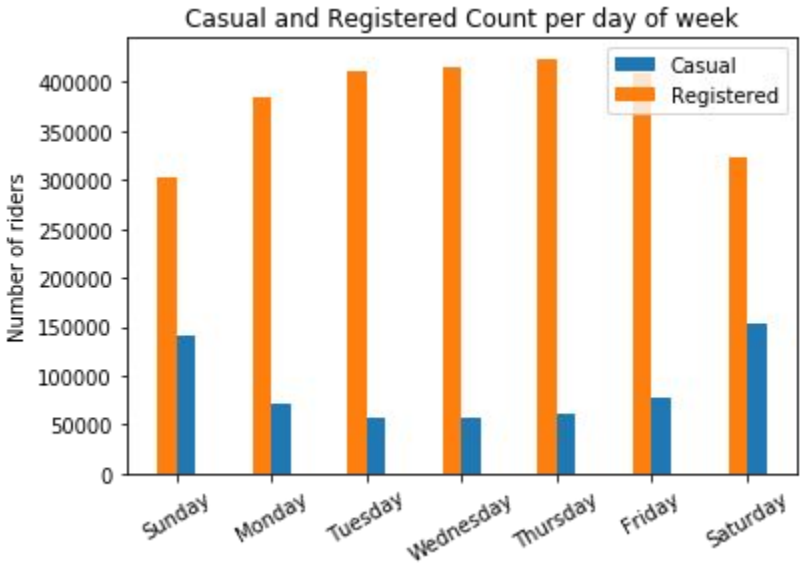
Problem

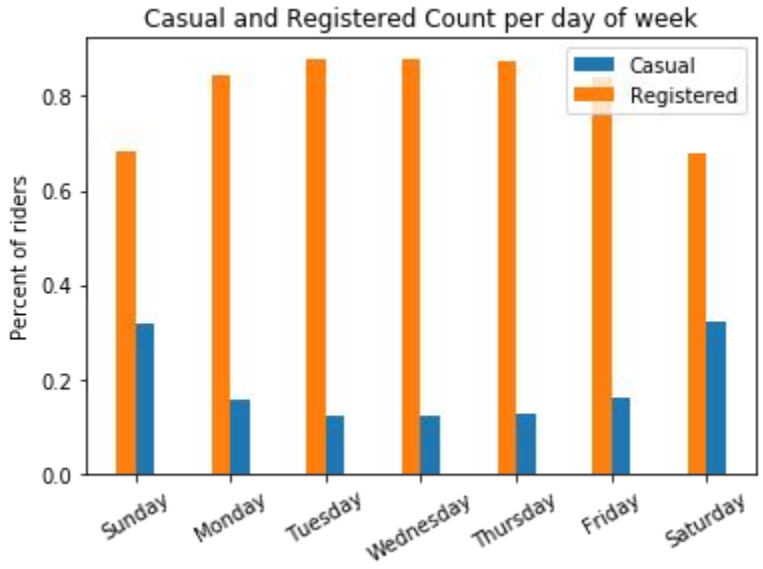
Want to predict the number of riders on the current day given the date and weather conditions.

Why?

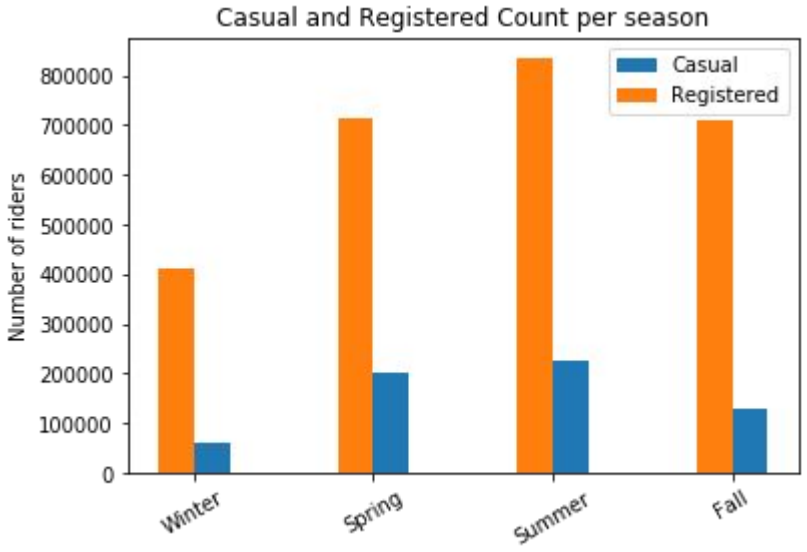
- Bike sharing companies
- Promotional offers
- Maintenance

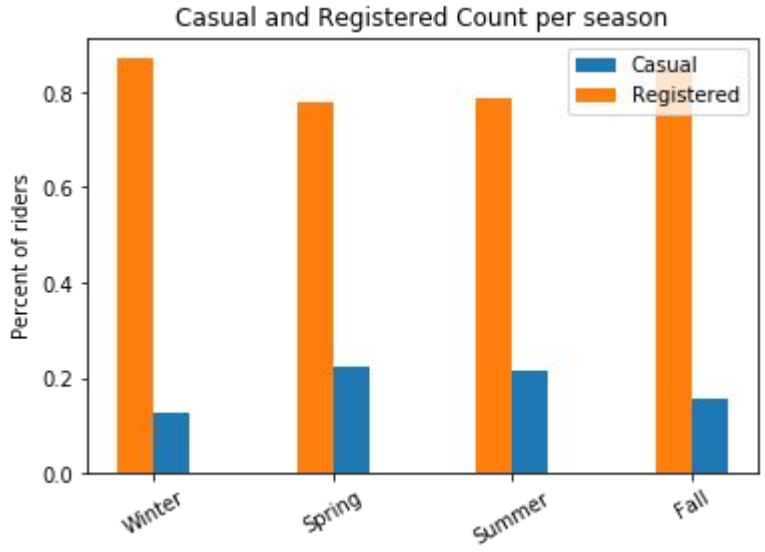
Data Exploration - Day of week



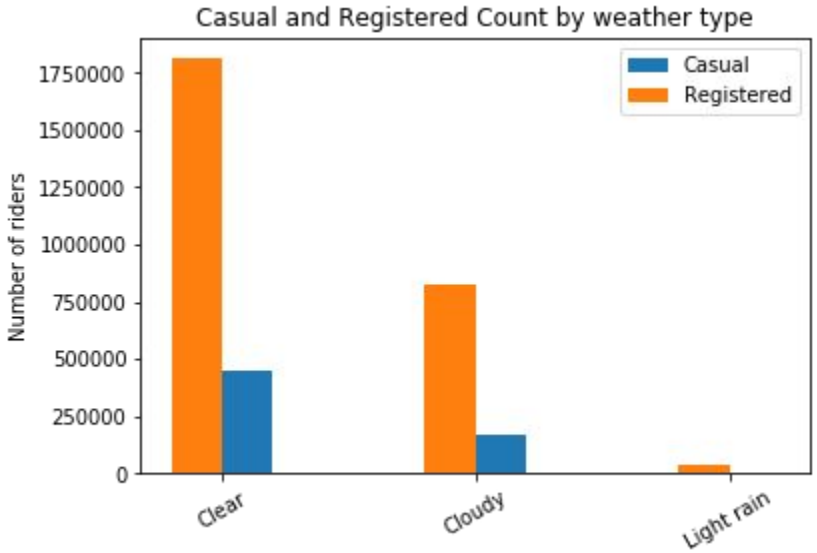


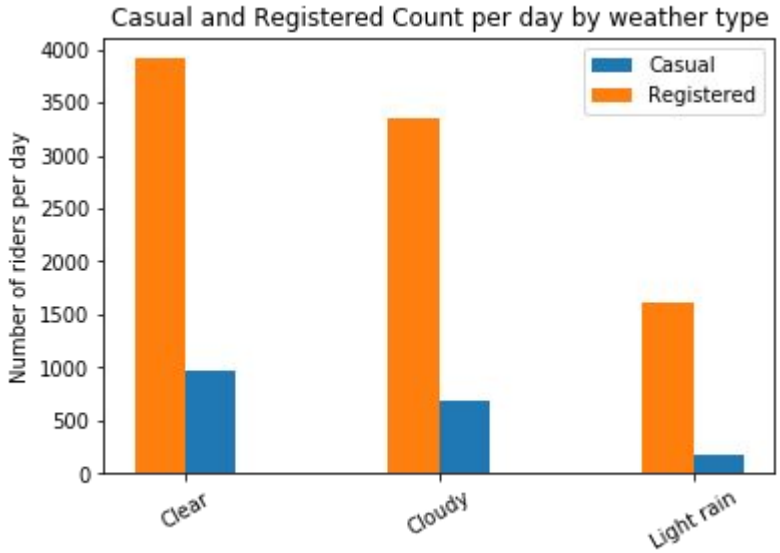
Data Exploration - Seasons





Data Exploration - Weathersit





Potential problems with data

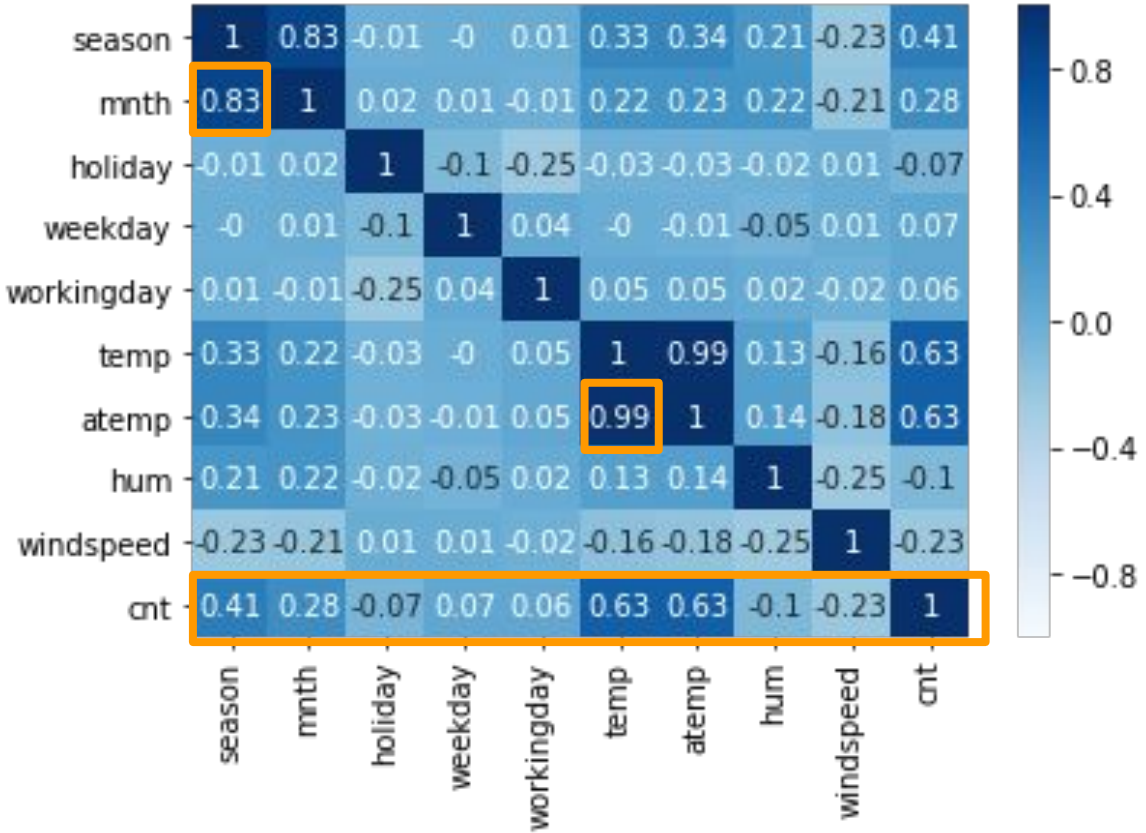
season : season (1:springer, 2:summer, 3:fall, 4:winter)

- Description is wrong, it’s actually (1:winter, 2:spring, 3:summer, 4:fall)

weathersit : (1:clear, 2:cloudy, 3:light rain, 4:heavy rain)

- No 4’s in data, seems wrong?
  - Hurricane Sandy
    - 2012-10-29
    - 22 riders, by far lowest, second lowest is 431
    - Hour data only has 3 4’s as well, typo?

Correlation



Next steps

Feature Engineering

Construct model, analyze performance, iterate on result

Find other outside sources of data to improve results