Bike Rental Analysis

Sam Skinner

12/3/2016

### Executive Summary

Bike Rentals are a growing industry but one that faces both weather-based and seasonal variations in demand. It is important to quantify how those those variations affect demand as demand, in turn, affects the number of employees needed as well as the number of bikes that need to be available. Significant mis-matches in bike availability or numbers of workers can result in missing rental opportunities (no bike available or insufficient employees for timely rental) or conversely, extra expenses to the rental company in terms of wages or bike purchases.

Using a publicly available dataset (described below), I attempt to answer two basic questions:

Q1: What are the factors that determine demand?

Q2: Do these factors vary by season?

The dataset consists of 17379 rental records collected over a two year period (2011, 2012) in Washington, D.C.. The records have been aggregated into 731 days for this analysis because each day is the determinant of the number bikes and employees needed.

Data attributes are both categorical and numerical and the statistical analyses vary accordingly.

##### The Data Source

The data are taken from the [UCI Machine Learning Repository](url:%20http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset). The original data are reported and analyized in a paper by [Fanaee, T and Gama, J.](url:%20http://link.springer.com/article/10.1007%2Fs13748-013-0040-3)

### Viewing the Data

As noted previously, the data consist of 731 day records with 15 variables per record plus a record number. Below is a list of the data attributes, including the variable name the type of data and a sample of the first 10 records for that attribute:

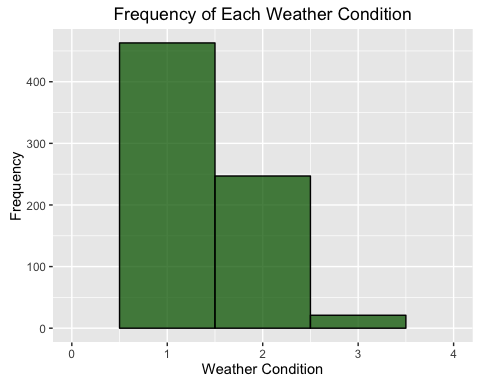
day <- read.csv('day.csv')  
str(day)

## 'data.frame': 731 obs. of 16 variables:  
## $ instance : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ date : Factor w/ 731 levels "1/1/11","1/1/12",..: 1 23 45 51 53 55 57 59 61 3 ...  
## $ season : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ year : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ month : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ holiday : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ weekday : int 6 0 1 2 3 4 5 6 0 1 ...  
## $ workingday: int 0 0 1 1 1 1 1 0 0 1 ...  
## $ conditions: int 2 2 1 1 1 1 2 2 1 1 ...  
## $ temp : num 0.344 0.363 0.196 0.2 0.227 ...  
## $ felt\_temp : num 0.364 0.354 0.189 0.212 0.229 ...  
## $ hum : num 0.806 0.696 0.437 0.59 0.437 ...  
## $ windspeed : num 0.16 0.249 0.248 0.16 0.187 ...  
## $ casual : int 331 131 120 108 82 88 148 68 54 41 ...  
## $ registered: int 654 670 1229 1454 1518 1518 1362 891 768 1280 ...  
## $ count : int 985 801 1349 1562 1600 1606 1510 959 822 1321 ...

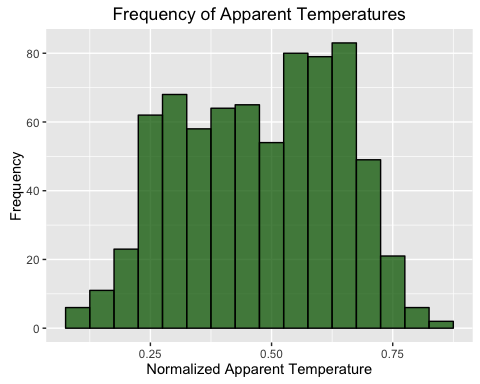
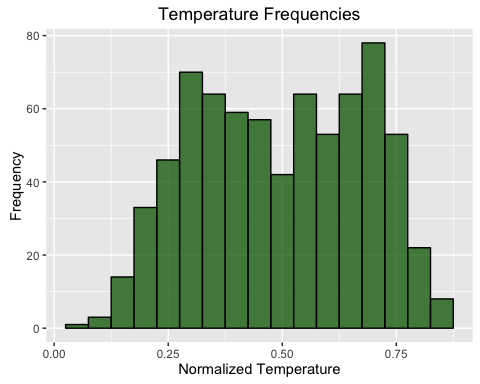
##### Description of Attributes

1. Instances: Record id number
2. date: Date (Factor w/ 731 levels "1/1/11","1/1/12",...)
3. season: Season (1:spring, 2:summer, 3:fall, 4:winter)
4. yr: Year (0:2011, 1:2012)
5. month: Month (1 to 12)
6. holiday: Holiday? (0:no, 1:yes)
7. weekday: Day of Week. Numbered 1-7, beginning with Sunday.
8. workingday: WorkingDay? (0:no, 1:yes)
9. conditions: Weather Conditions: (1: Clear or Partly cloudy, 2: Mist with Few Clouds up to Cloudy, 3: Light Snow or Light Rain with Scattered Clouds/Cloudy/Thunderstorm, 4: Heavy Rain, Hail or Thunderstorm or Mist or Snow + Fog)
10. temp: Normalized temperature in Celsius. The values are divided to 41 (max)
11. felt\_temp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
12. hum: Normalized humidity. The values are divided to 100 (max)
13. windspeed: Normalized wind speed. The values are divided to 67 (max)
14. casual: Count of casual users
15. registered: Count of registered users
16. count: Count of total bike rentals including both casual and registered

#### Visuals

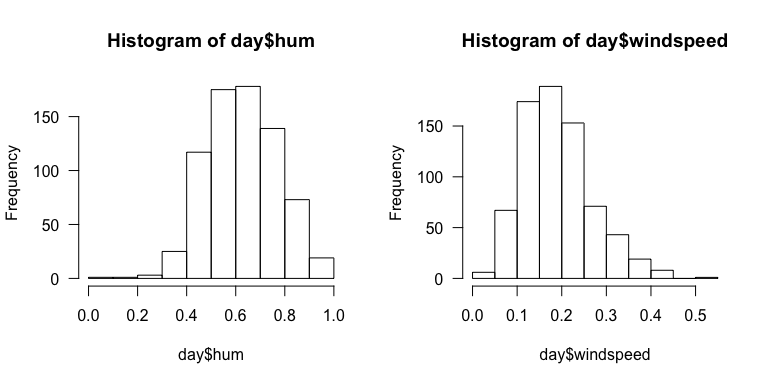


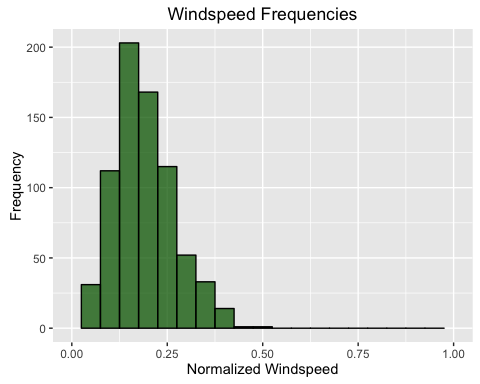
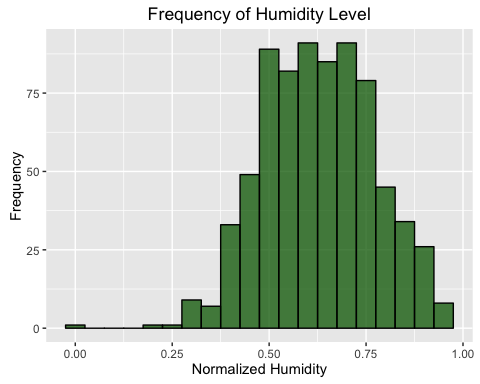
#### Notice that there are no "4" values and very few "3" values (n= 21). In other words, we generally have sunny or partially cloudy days.



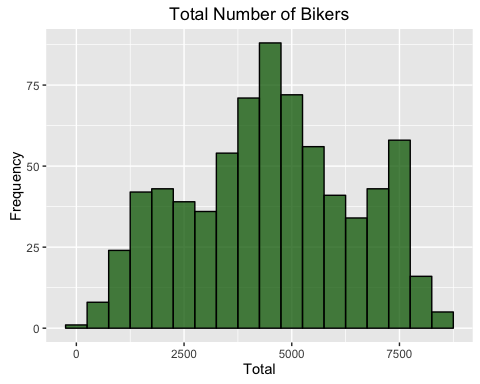
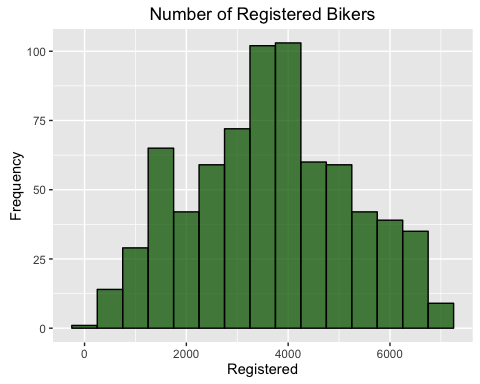
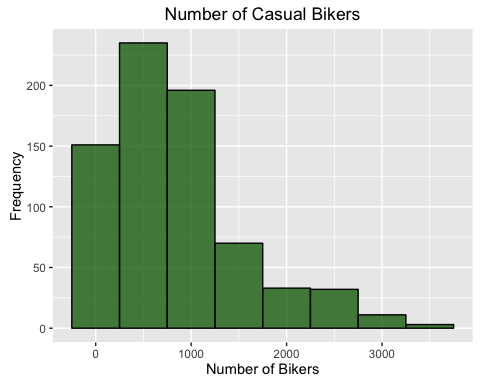
#### Both distributions are platykurtic.

par(mfrow=c(1,2), las=1)  
for(i in 1:2) {  
hist(day$hum)  
hist(day$windspeed)  
}

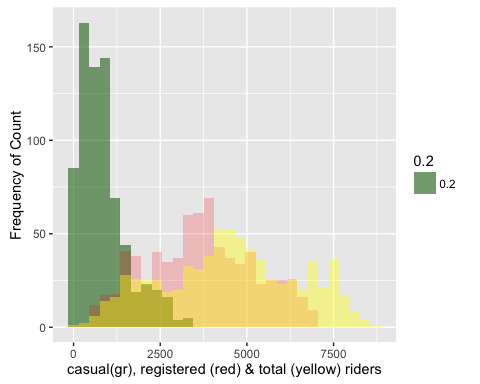




#### The "zero" humidity day appears to be an error, perhaps a day of no data. Notice that the two distributions are slightly skewed, humidity to the right, windspeed to the left.

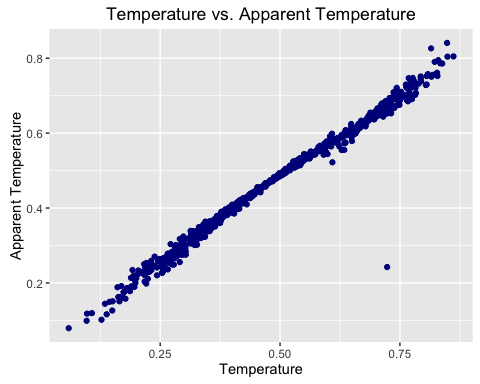


## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



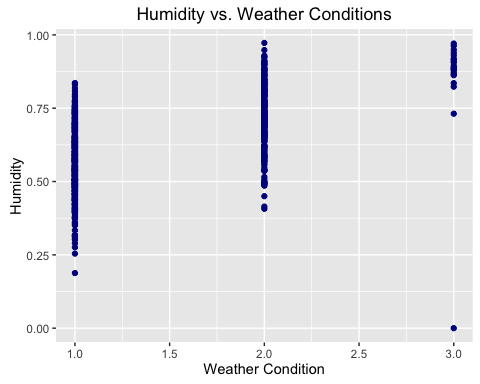
### Reducing the Dataset

The current dataset contains variables that are irrelevant to our purposes. We don't need the record numbers (though they were useful above in identifying and evaluating possible outliers); nor do we need the specific dates because we will confine our analysis to month and season. Additionally, actual temperature and felt ("apparent") temperature are highly correlated (see plot correlation coefficient = 0.99) so I eliminated the latter.



Next I looked at the relationships between "conditions" and the individual weather measurements. I used Spearman's rank correlation tests because "conditions" is an ordinal variable, not continuous. The correlation between condition and both temperature and windspeed were non-significant (not shown). However, there was a significant relationship between humidity and condition, with a Spearman's rho of 0.597 at the p < 2.2e-16 significance level (see below). This indicates that humidity is a pretty good substitute for "condition".

Notice the one extreme outlier. It shows zero hunidity on a day where the weather condition is "light rain or snow". Without this datum, the correlation would undoubtedly be higher, reinforcing the previous conclusion. [Additional testing of the power of "condition" for predicting numbers of bike riders also showed it having no better power than humidity, confirming the appropriateness of removing "condition".]



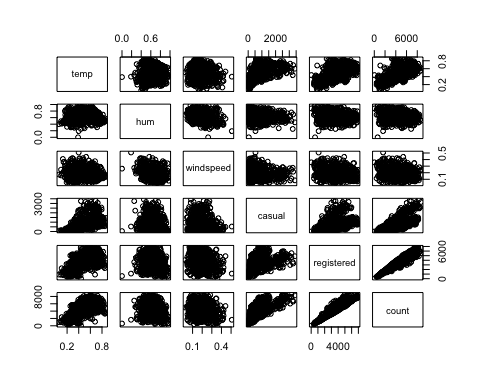
## Warning in cor.test.default(day$conditions, day$hum, method = "spearman"):  
## Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: day$conditions and day$hum  
## S = 26267000, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.596532

Finally, I also looked at possible relationships among other weather variables and these, in fact, show various levels of correlation, generally with significance levels well below 0.01. However, in each case the R-squared values are below 0.25 indicating that the relationship has low explanitory power. Consequently, none of the remaining variables were eliminated.

Having reduced the number of variables by four, we can take a look at the summary of the non-categorical variables(below). The analyses to follow are based on this reduced set of variables.

## temp hum windspeed casual   
## Min. :0.05913 Min. :0.0000 Min. :0.02239 Min. : 2.0   
## 1st Qu.:0.33708 1st Qu.:0.5200 1st Qu.:0.13495 1st Qu.: 315.5   
## Median :0.49833 Median :0.6267 Median :0.18097 Median : 713.0   
## Mean :0.49538 Mean :0.6279 Mean :0.19049 Mean : 848.2   
## 3rd Qu.:0.65542 3rd Qu.:0.7302 3rd Qu.:0.23321 3rd Qu.:1096.0   
## Max. :0.86167 Max. :0.9725 Max. :0.50746 Max. :3410.0   
## registered count   
## Min. : 20 Min. : 22   
## 1st Qu.:2497 1st Qu.:3152   
## Median :3662 Median :4548   
## Mean :3656 Mean :4504   
## 3rd Qu.:4776 3rd Qu.:5956   
## Max. :6946 Max. :8714



### Univariate Models

#### Weather Variables on Number of Casual or Registered Riders

### Multivariate Models

### The Preferred Predictive Model

### What Next?

### About the Author



[Sam Skinner](https://www.linkedin.com/in/samskinnerphd) is a rambling rube in R-land, hoping to hitch a ride with fellow travelers through the R-landscape.



© Sam Skinner, 2016.