Data 583: Project Proposal

Ricky Heinrich & Vimaljeet Singh

2023-02-27

About Data Collection

The dataset we have chosen to analyze is Hadi Fanaee-T's Bike Sharing Dataset, from the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, accessed in the UCI Machine Learning Repository. This dataset combines the Trip History Data for the years of 2011 and 2012 of 'Capital Bikeshare', which is metro Washington DC's bikeshare service, with weather data and the holiday schedule. The data consists of an aggregated count of 'rides' by hour (hour.csv, 17379 x 17) and by day (day.csv, 731 x 16).

Capital Bikeshare's data comes from their electronic systems. An 'observation' is collected for every 'trip' a bike takes. It has been processed to include only 'valid' trips from the user base, removing service trips and short trips (less than 1 minute), attribute to potential false starts.

The dataset source referencing the site https://ca.freemeteo.com/ as their weather data source. We presume they scraped or used an API to access historical weather data for the DC region, and that the result is an hourly average.

Their reference link to the Holiday Schedule data is broken, but we assume the current link is (https://dchr.dc.gov/page/holiday-schedules). The site has since updated to show the 2023 holiday schedule, but looking back on the WayBack Machine with the original link we see what the page looked like in 2013, when the data was donated. We presume again that the data was either webscraped or entered manually.

Statistical Description

Data Types and Structure

Most of the data is numerical in nature, apart from the *dteday* column, which records the date in a date format. This date is separated further in year, month, and hour columns, which are factored into discrete numeric variables, similarly to the *weekday* variable. The *year*, *holiday*, and *workinday* variables are boolean variables. For the *year*, a '0' represents 2011, and a '1' 2012.

The *weathersit* variable is categorical, where the conditions were classified into four. The data can be taken as ordinal, with '1' being the most 'pleasant' weather, and 4 the least. From the original data source description, 1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog.

The rest of the weather related variables, *temp*, *atemp*, *hum* and *windspeed* have been normalized. From the original data source description: *temp* : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale); *atemp*: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale); *hum*: Normalized humidity. The values are divided to 100 (max); *windspeed*: Normalized wind speed. The values are divided to 67 (max).

Finally, *casual*, *registered*, and *cnt* are counts of bikes rented during each 'hour', corresponding to the count of casual users, registered users, and the sum of both.

Distributions

We've generated "histograms" of each variable and analyzed them to infer their distribution. Plots are available in "Plot Appendix" at the end of the report.

We are expecting that the counts of observations for each year, weekday and hour are uniform across categories: each year should have the same number of hours, same as weekday and hour. We suppose that there are less 'observations' for the hours of 2,3,4,5 due to some days containing no rides during that hour. Otherwise, every other hour in the two year period contains at least one ride.

The season, month, holiday, and workingday 'counts' are not interesting on their own: they are constant values for a given period of time.

In the 'weathersit' plot, we see that a majority of cases are classified as '1', and decreasing counts as the weather gets less pleasant. This tells us that the weather is pleasant most of the time in DC.

The other weather variables are normalized. We may reverse engineer the raw values, given the normalization formula, for final analysis. In the 'temp' plot, we see what looks like a symmetric bimodal distribution. We know that the minimum is -8, and the maximum is +39 Celsius. 'atemp', which is the feeling temperature, shows more of a flattened peak, although there is one value that is recorded about doubly more often than any other. Similarly there's a few troughs. Generally it seems symmetric. We would have expected a more continuous distribution, so there might be something going on in regards to rounding or collection of data. Similarly, I would have expected more continuous data for the humidity records, as well as the windspeed. We see how the data is not exactly normally distributed from the qqplots.

For the count of casual users during a given hour, we see what looks like a steep exponential decline. Intuitively, less users during a given hour happen a lot more often then a lot of users. The distribution of registered users sees less of an extreme drop, with a less steep decline from 50 counts on of individual hours observing a minimum of about 75 rides on. We see that there is a lot more counts of casual users however (peaking >1500 vs >300), so on the same y-scale the distributions might look different. More investigation is needed in that regard.

Underlying scientific processes that may affect the data

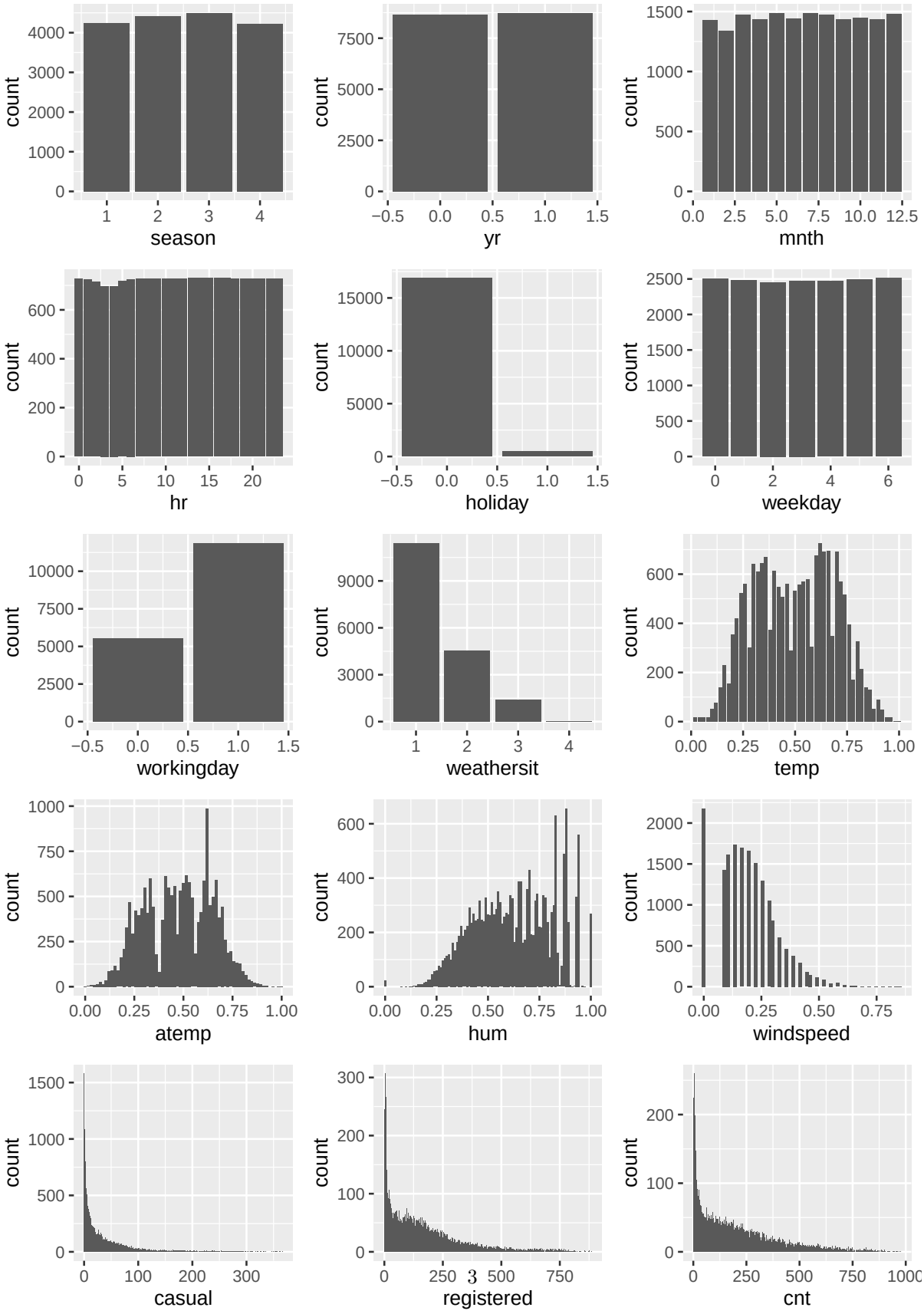
Underlying scientific processes that might have affected the data would be the occurrence of natural events like Hurricane Sandy on Oct 30, 2012. So, at instances where data might seem aberrant, it might be because the count of bike users was affected by these natural events or anomalies.

Scientific questions

- Can we detect any anomalies or events that affect the rental count, such as holidays or weather events, using the data set?
- Can we predict the future demand for bike rentals based on past data and environmental factors?
- How does the demand for bike rentals differ between registered and casual users?
- What is the impact of seasonal changes on bike rental demand?
- How does the rental count of bikes vary with seasonal and environmental factors such as weather conditions, temperature, humidity, and wind speed?
- Can we identify any patterns in the hourly rental counts, such as spikes during commuting hours?
- What affects the number of bike rides the most?

Plot Appendix

Distribution plots



QQ plots for ‘temp’, ‘atemp’, ‘hum’, and ‘windspeed’