EDA

Ricky Heinrich

2023-03-10

Workflow for current and next step in the project

To determine which model would fit this data, we need to consider several factors such as the nature of the problem, the available data, and the type of output that we are trying to predict.

First, we need to define the problem and the goal of the model. In this case, we are trying to **predict the number of bikes rented per hour (cnt)**, given the various features such as weather, time of day, and other factors. This is a regression problem, as we are trying to predict a discrete variable.

Second, we need to examine the available data and determine if there are any missing values, outliers, or other anomalies. If the data is incomplete, we may need to consider techniques such as imputation or data cleaning to address these issues.

Third, we need to select appropriate features for the model. Some features, such as season or weather, may have a strong correlation with the number of bikes rented, while others may not be as important. Feature selection can be done using techniques such as correlation analysis or principal component analysis.

Finally, we can select a model that is appropriate for the problem at hand. Some popular models for regression problems include linear regression, decision trees, random forests, and neural networks. We can use techniques such as cross-validation to evaluate different models and select the one that performs best on the data.

Statistically Descriptive Analysis of the Dataset

the number of variables, variables types, summary statistics, graphs of data

Date related variables The dataset contains 17 variables, where the first column is an index. 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, where Sunday is 0. The *year*, *holiday*, and *workinday* variables are boolean variables. For the *year*, a '0' represents 2011, and a '1' 2012.

We were expecting that the counts of observations for each year, weekday and hour are uniform across categories, as each should have the same number of hours. We suppose that there is a dip in 'observations' for the hours of 2,3,4,5 due to some days containing no rides during that hour. The count of hours with rides is slightly lower in 2011 than 2012, and we see a small concave curve with Sunday and Saturday on both ends and where there minimum count is on Tuesday.

The distributions of counts for season, month, holiday, and workingday are not interesting on their own, as each category is not meant to have the same number of hours (February has 72 less hours total than March since its got 3 days less). We cannot infer if the dips are due to hours containing no rides or just how the categories are set.

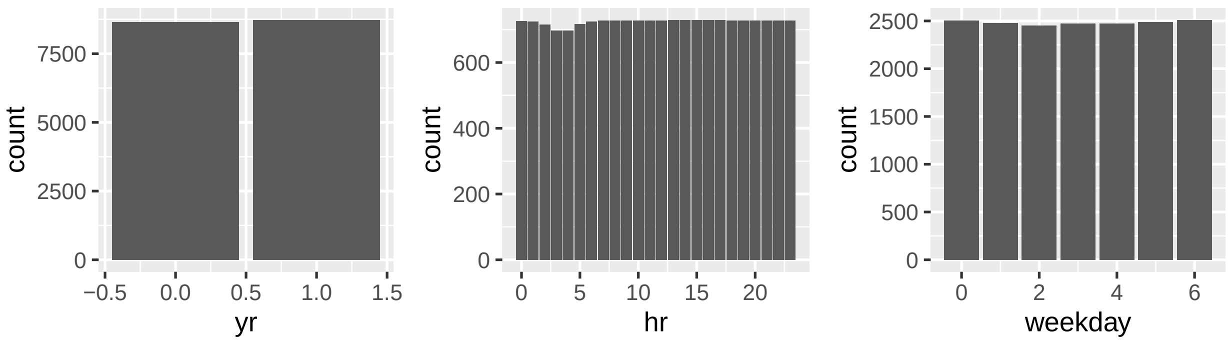


Figure 1: Count of rows per Year, Hour, and Weekday

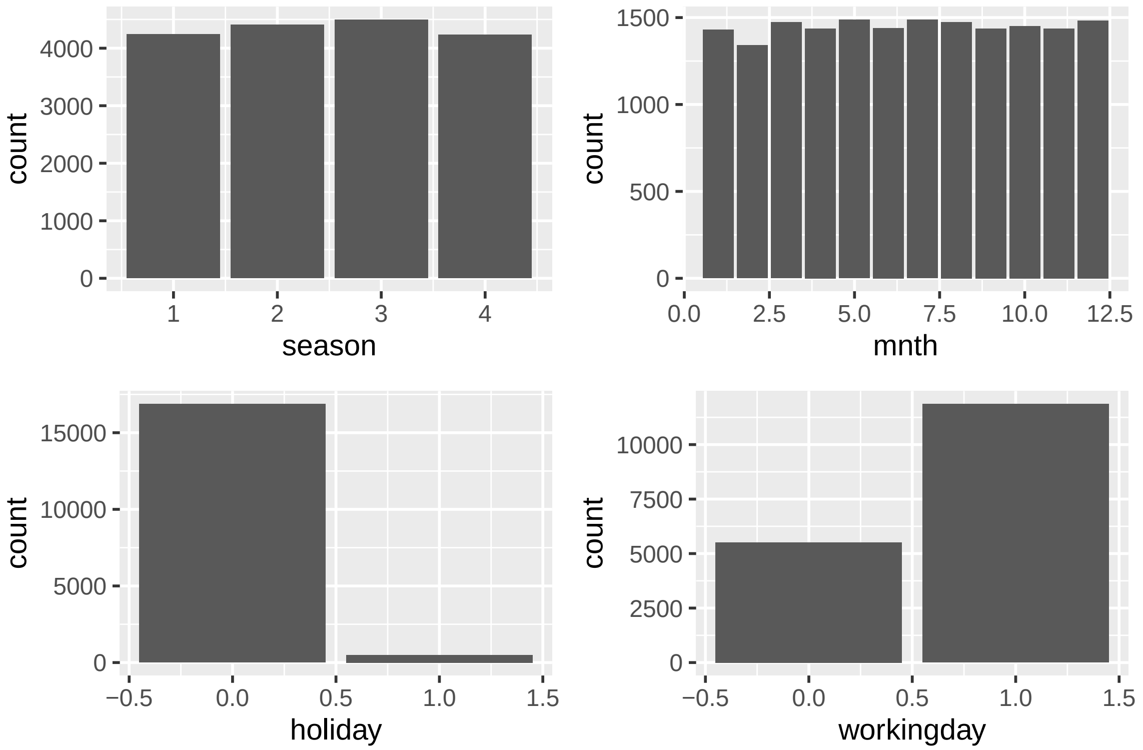


Figure 2: Count of rows per Season, Month, Holiday, and Workingday

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.

We see that a majority of cases are classified as '1', and decreasing counts as the weather gets less pleasant. We know the data doesn't contain an exhaustive list of all hours, but assuming the effect of missing hours is not great, this tells us that the weather is pleasant most of the time in DC.

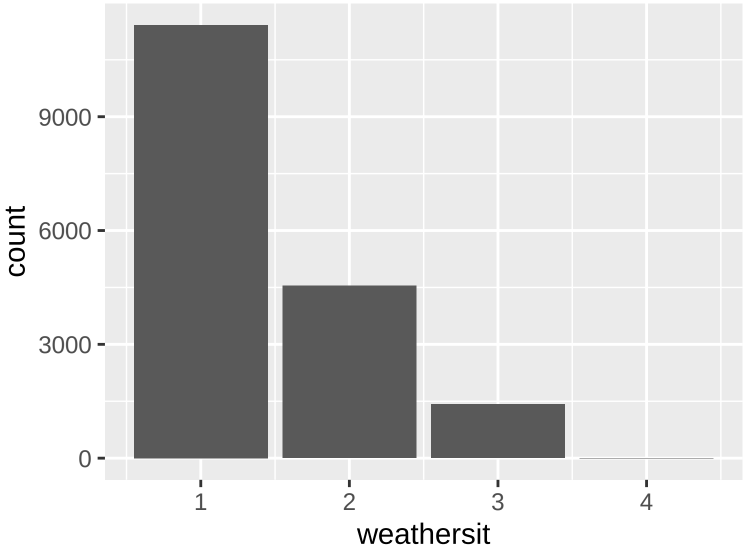


Figure 3: Count of rows per weather category

The rest of the weather related variables, *temp*, *atemp*, *hum* and *windspeed* have been scaled. We've transformed the data in new columns to get back original values to help make charts interpretable. 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, but 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, we would have expected more continuous data for the humidity records, as well as the windspeed.

```
summary(df[18:21])
```

```
##     rawtemp         rawatemp          rawhum       rawwindspeed
##  Min.   :-7.06   Min.   :-16.000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.: 7.98   1st Qu.:  5.998   1st Qu.: 48.00   1st Qu.: 7.002
##  Median :15.50   Median : 15.997   Median : 63.00   Median :12.998
##  Mean   :15.36   Mean   : 15.401   Mean   : 62.72   Mean   :12.737
##  3rd Qu.:23.02   3rd Qu.: 24.999   3rd Qu.: 78.00   3rd Qu.:16.998
##  Max.   :39.00   Max.   : 50.000   Max.   :100.00   Max.   :56.997
```

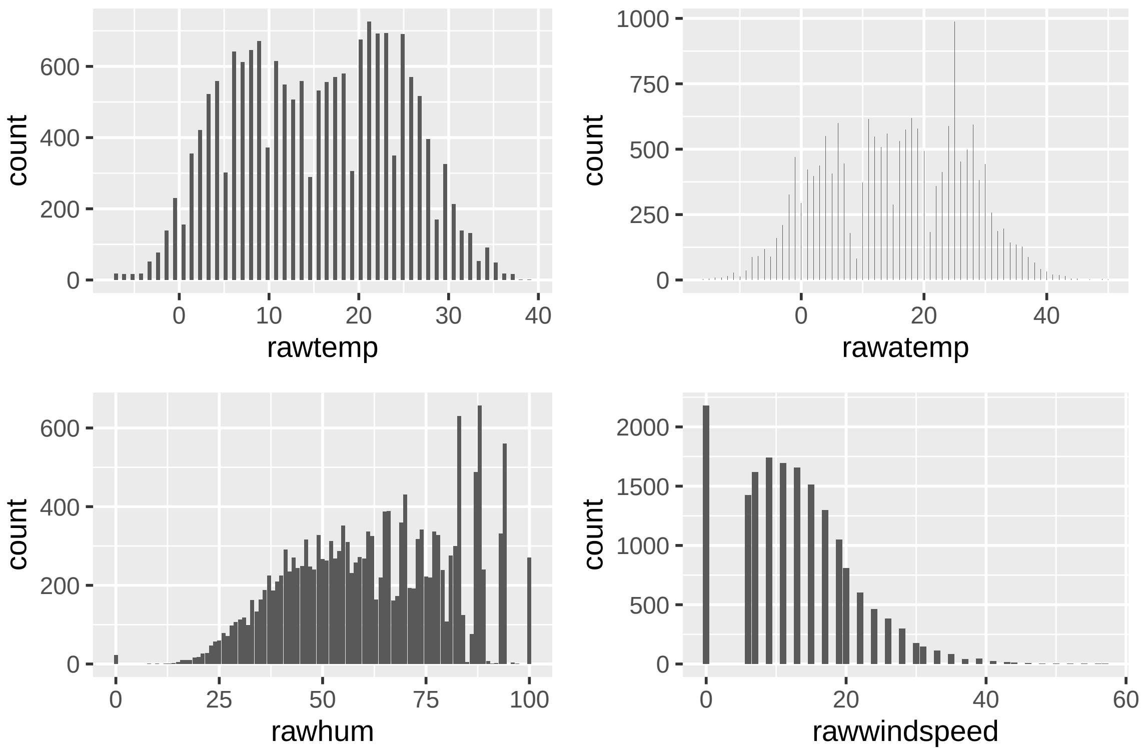


Figure 4: Count of rows per temp, atemp, hum, and windspeed

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. 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.

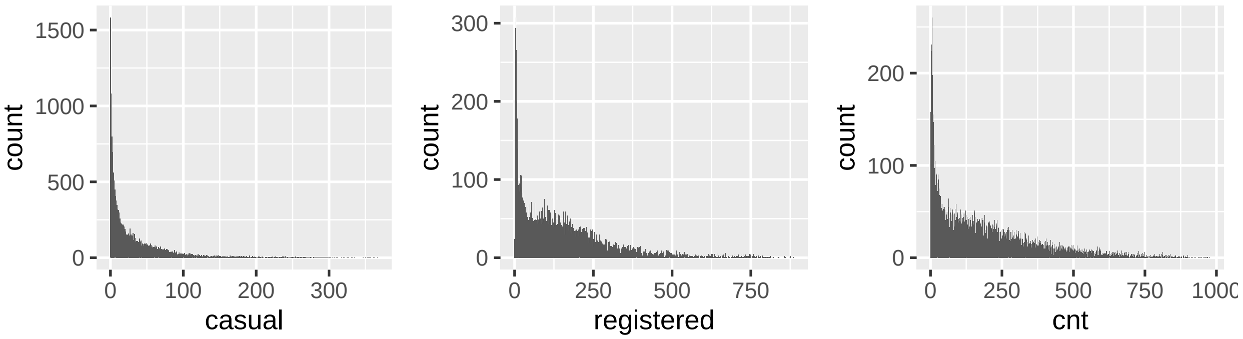


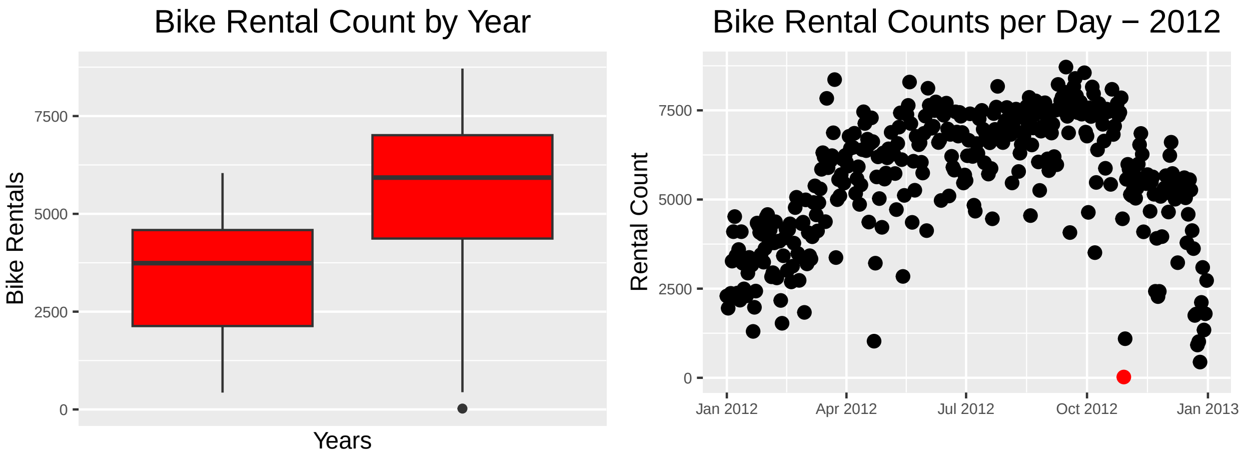
Figure 5: Count of rows per temp, atemp, hum, and windspeed

```
summary(df[15:17])
```

```
##       casual          registered         cnt
##  Min.   :  0.00   Min.   :  0.0   Min.   :  1.0
##  1st Qu.:  4.00   1st Qu.: 34.0   1st Qu.: 40.0
##  Median : 17.00   Median :115.0   Median :142.0
##  Mean   : 35.68   Mean   :153.8   Mean   :189.5
##  3rd Qu.: 48.00   3rd Qu.:220.0   3rd Qu.:281.0
##  Max.   :367.00   Max.   :886.0   Max.   :977.0
```

```
# trying to build a box plot with all of these on one plot ..
#ggplot(df, aes())
```

To see if Bike Rental Count is different for both years



- The number of bikes rented out has increased from 2011 in 2012.
- Also, we notice the outlier in 2012 which can be seen as a dot towards the bottom of the plot, plotted in red in the scatterplot.

Let us explore this further:

```
myday_2012 <- subset(day_sum, yr == 2012)
mydata = myday_2012[myday_2012$cnt == min(myday_2012$cnt), c("cnt","dteday")]
cbind(yearly_mean = mean(myday_2012$cnt), mydata)
```

```
##     yearly_mean cnt     dteday
## 668    5599.934  22 2012-10-29
```

We see that the number of bikes rented out on October 29, 2012 was underwhelmingly lower than the yearly average for 2012 making it an outlier. We investigated further to see why that is and found the following:

!

It was because of this natural calamity that the number of bike rental counts was super low. We have also successfully used this data set to find a natural calamity.

Check if 'weathersit' and 'holiday' variables are important using ANOVA

```
lm <- lm(cnt ~ weathersit + holiday, data = df)
anova (lm)
```

```
## Analysis of Variance Table
##
## Response: cnt
##               Df    Sum Sq  Mean Sq F value    Pr(>F)
## weathersit     1  11598301 11598301 360.183 < 2.2e-16 ***
## holiday        1    636252   636252  19.759 8.841e-06 ***
## Residuals  17376 559527039    32201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F values greater than 1, and pvalues less than 0.05 show that both these variables are significant in predicting count of bike rentals.

average bike rental per month, per hour, per weekday

what questions we will answer: how will try to answer

- what is most influencal variable: do that random forest gni index thing, which variable has most different MSE when removed ?
- can we predict number of bikes given set of conditions : regression, could try non-parametric even tho you said linear regression is perfect fit lol
- maybe we could try to cluster registered vs not registered. actually probably can't since its aggregated, this would be something from og og data

I think the normality plots about the weather data are not that interesting and I don't see the purpose of including them

I like bike rental counts vs temperature, still think the colour scale is not the most straighforward so maybe we can come up with something more