HAMLINE UNIVERSITY

Data Mining: Problem Set 2

Xander C*

09-25-2023

```
  knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE, include = TRUE)
# Load libraries
  library(readr)
  library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
  library(tidyr)
  library(e1071)
  library(ggplot2)
  library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
  library(caret)
```

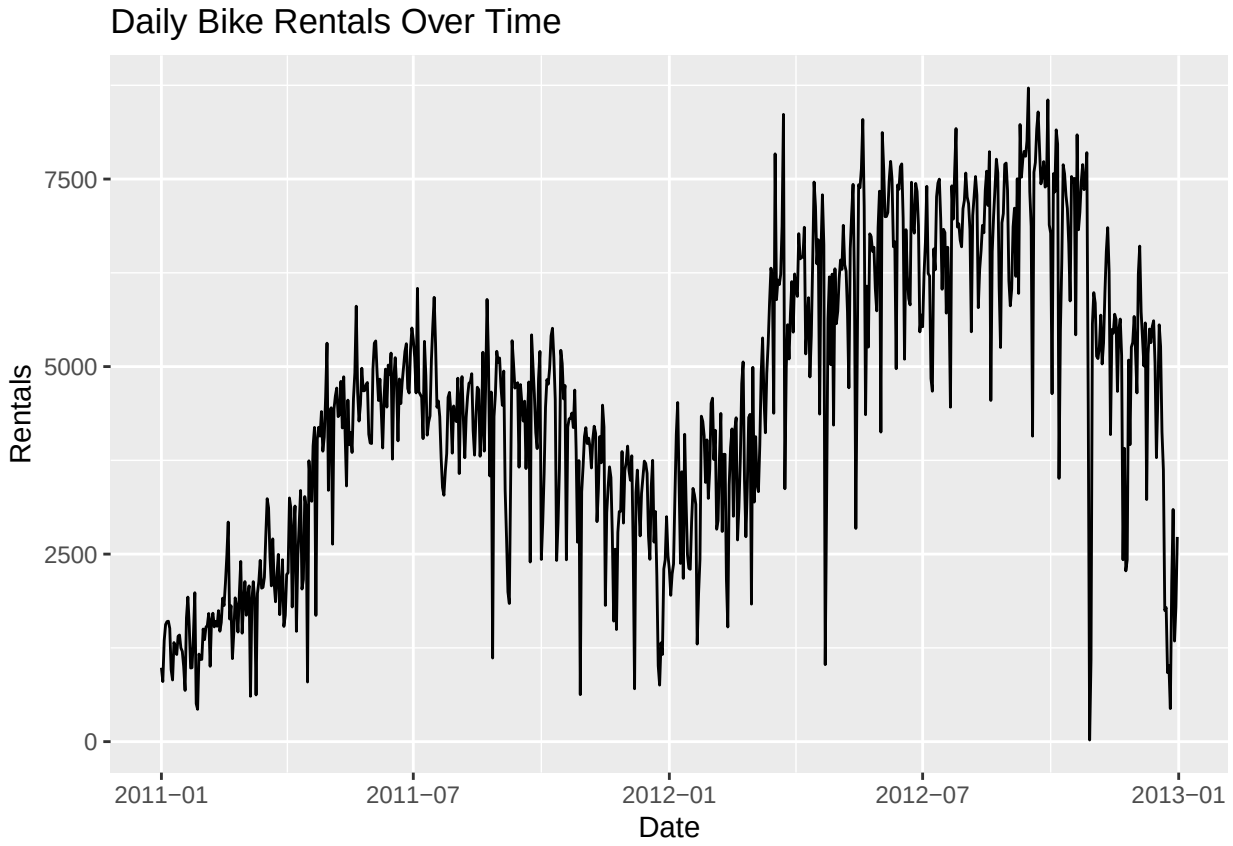
```
## Loading required package: lattice
```

```
  library(caretEnsemble)
```

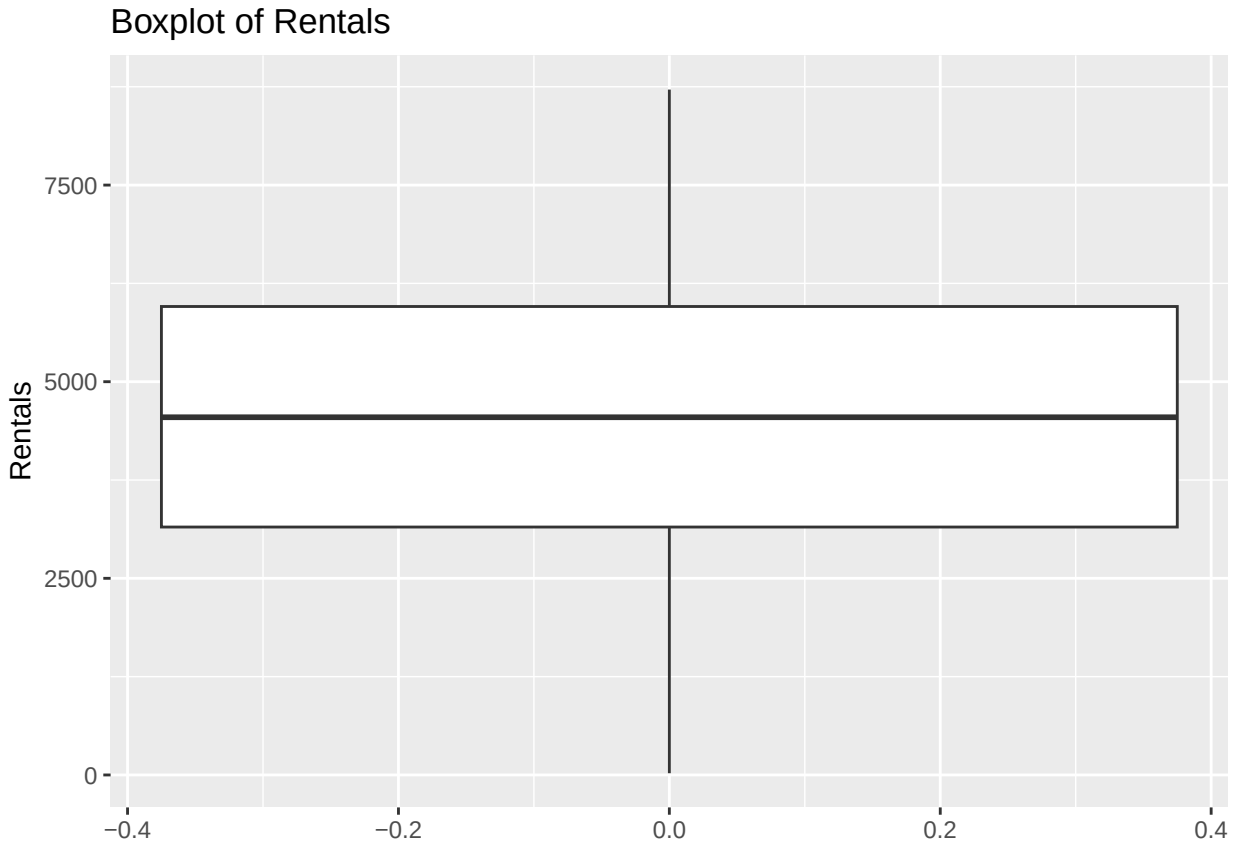
```
##
## Attaching package: 'caretEnsemble'
```

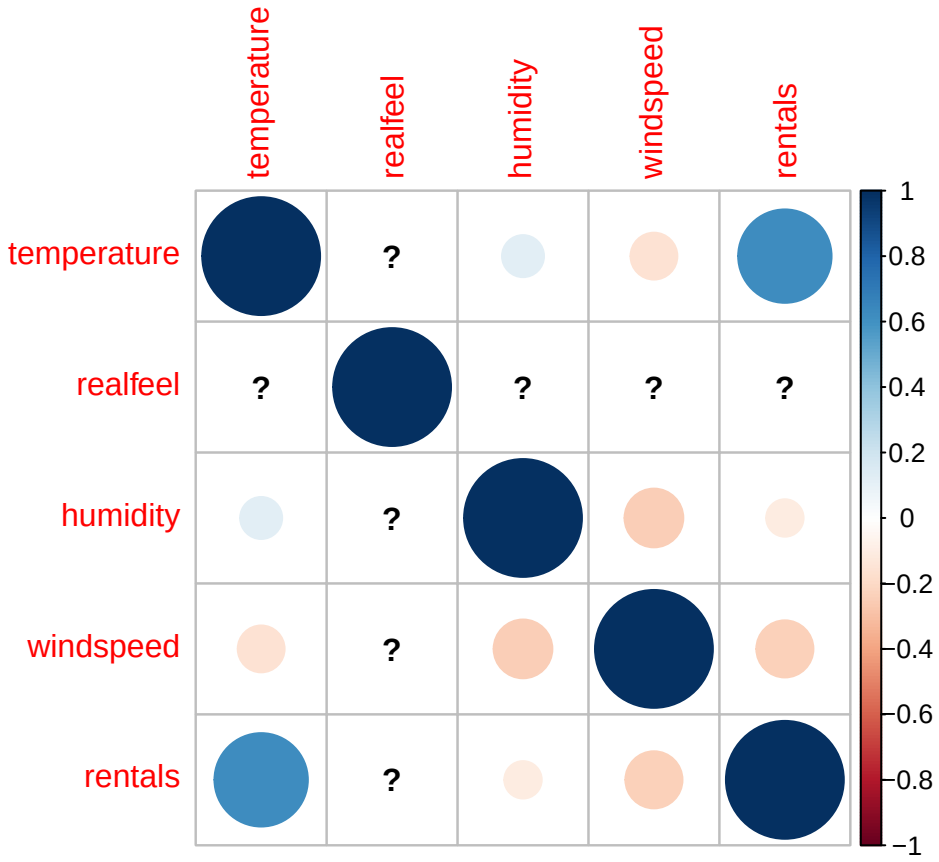
*Email achapman03@hamline.edu. Position Student

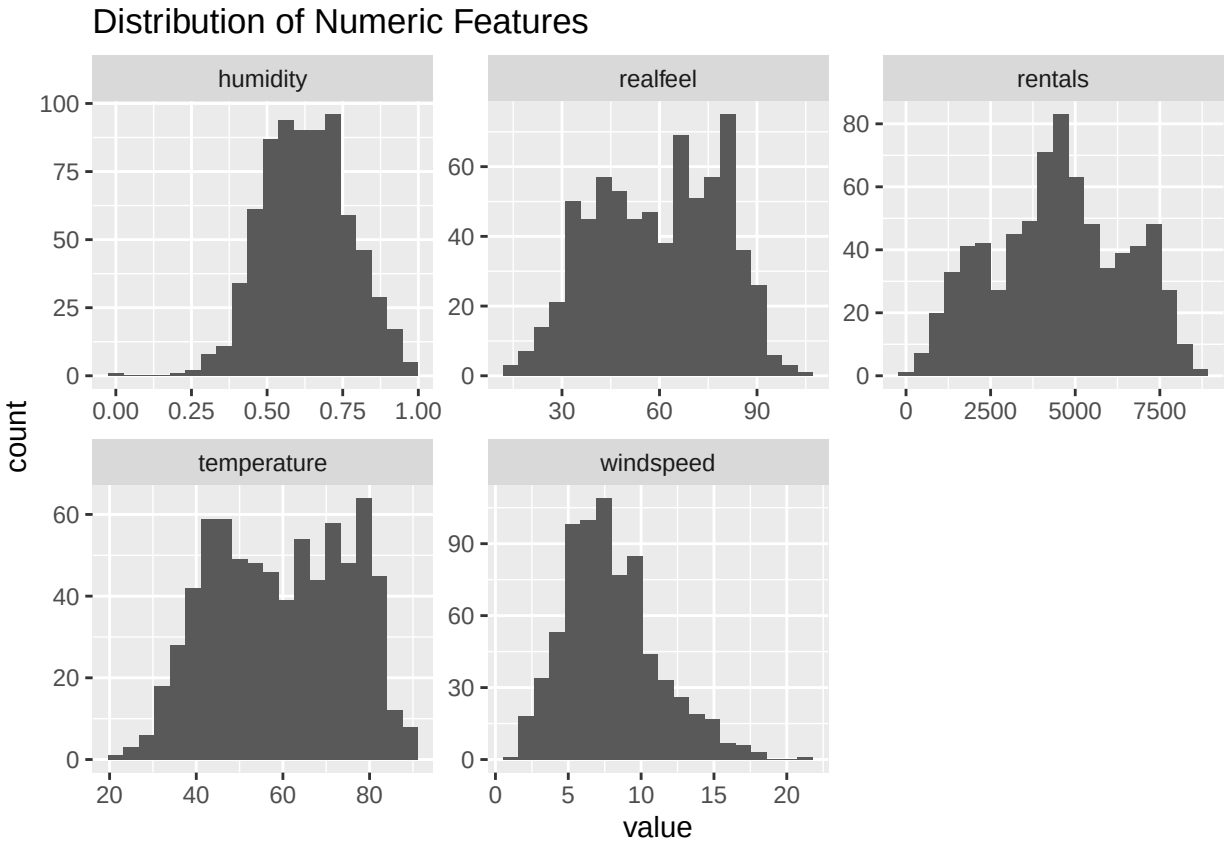
```
## The following object is masked from 'package:ggplot2':
##
##     autoplot

## Number of rows: 731

## Number of features: 10

## spc_tbl_ [731 x 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date       : Date[1:731], format: "2011-01-01" "2011-01-02" ...
##  $ season     : num [1:731] 1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday    : num [1:731] 0 0 0 0 0 0 0 0 0 0 ...
##  $ weekday    : num [1:731] 6 0 1 2 3 4 5 6 0 1 ...
##  $ weather    : num [1:731] 2 2 1 1 1 1 2 2 1 1 ...
##  $ temperature: num [1:731] 46.7 48.4 34.2 34.5 36.8 ...
##  $ realfeel   : num [1:731] 46.4 45.2 25.7 28.4 30.4 ...
##  $ humidity   : num [1:731] 0.806 0.696 0.437 0.59 0.437 ...
##  $ windspeed  : num [1:731] 6.68 10.35 10.34 6.67 7.78 ...
##  $ rentals    : num [1:731] 985 801 1349 1562 1600 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   season = col_double(),
##   ..   holiday = col_double(),
##   ..   weekday = col_double(),
##   ..   weather = col_double(),
##   ..   temperature = col_double(),
##   ..   realfeel = col_double(),
##   ..   humidity = col_double(),
##   ..   windspeed = col_double(),
##   ..   rentals = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

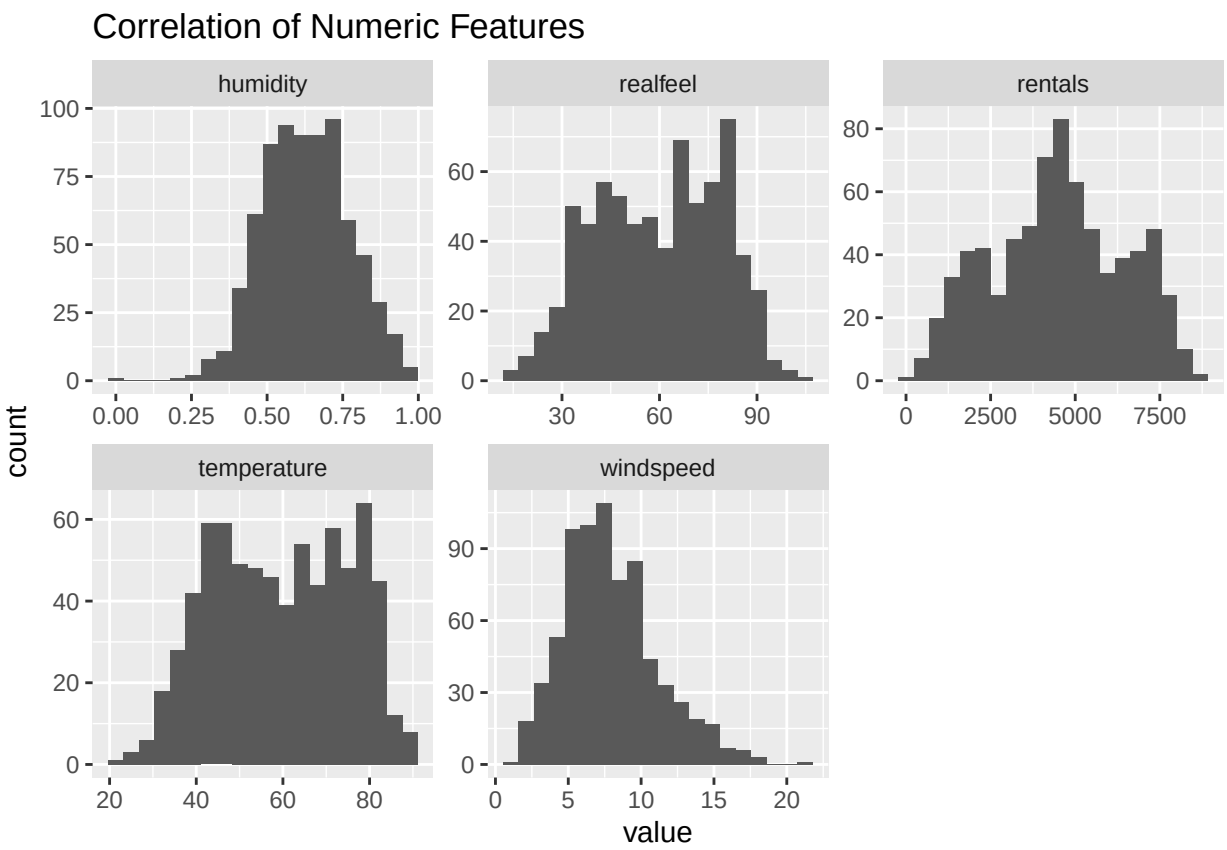


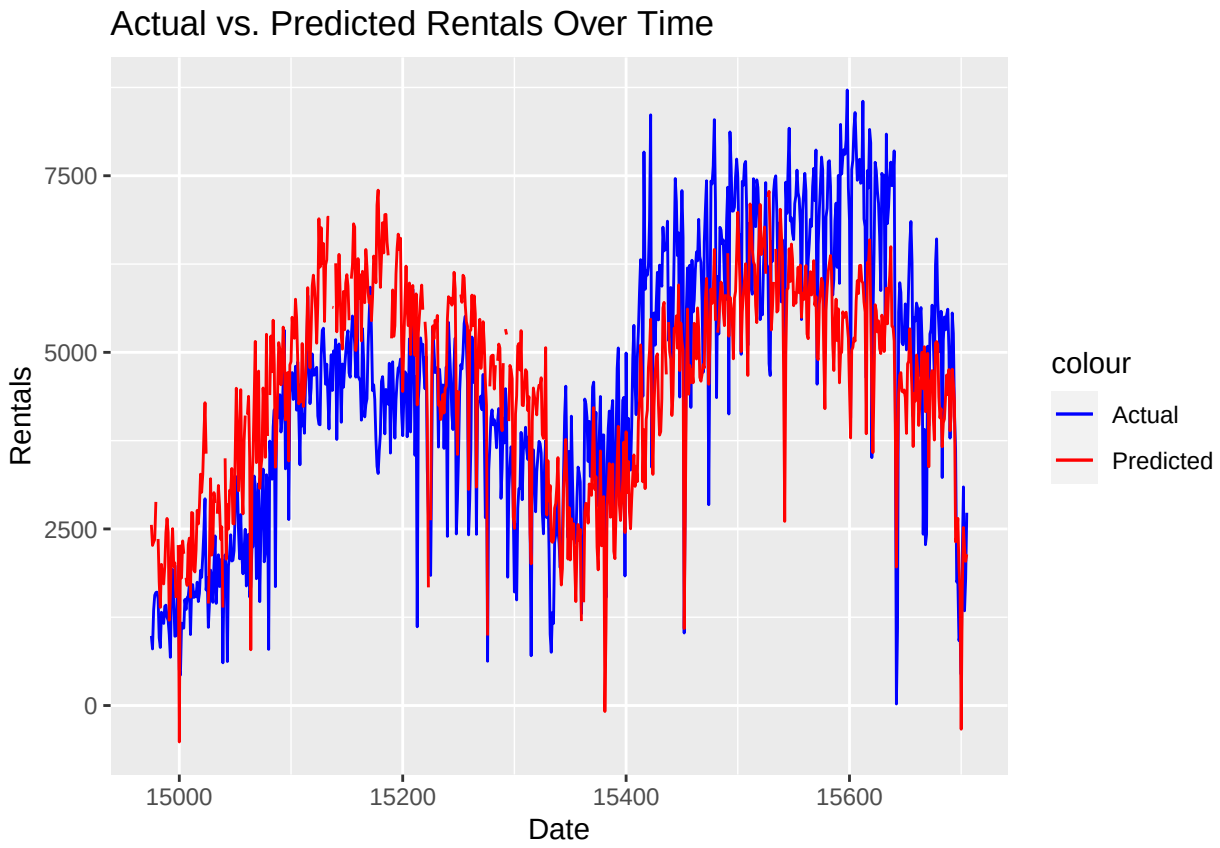

```
## Skewness of Rentals: -0.04715862
```


Boxplot of Rentals
Rentals
7500
5000
2500
0
−0.4
−0.2
0.0
0.2
0.4


temperature
realfeel
humidity
windspeed
rentals
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Distribution of Numeric Features
humidity
realfeel
rentals
temperature
windspeed
count
value


Correlation of Numeric Features
humidity
realfeel
rentals
temperature
windspeed
count
value



Discuss whether you think the features you have in the data make sense for learning to predict daily bike rentals.

I believe the features available in the data make sense for learning to predict daily bike rentals. We are able to compare multiple important factors such as weather, day of week, is it a holiday, or what the weather is to be able to predict the amount of bikes that will be rented.

Discuss what is means in this case to train or "fit" a model to the data you prepared.

Training a model in this case means comparing real life factors such as weather, day of week to the amount of rentals. This allows the model to learn the rental patterns depending on the day, season, or weather so that the model can predict what future rentals will look like depending on the season, day of the week, and holidays.

Discuss which preparations you did were required to make the learning algorithm work, and which were not strictly required, but maybe are a good idea.

The required preparations made to make the algorithm work were data type conversion, median imputation, normalization, dummy variables, and training the model. These preparations

are required for the algorithm to give us an answer to our question. If these had not been completed, the algorithm would either not give us an answer, or give us a useless answer. The best practice preparations were data visualization, correlation analysis, outlier detection, and model evaluation. These preparations allow for us to make the algorithm work for us. Visualizing, analysizing, outlier detection, and model evaluation, allow for us to tune the algorithm which can lead to better answers.