Homework 3 - BUSN 41204

Fernando Garcia, Paula Gaviria, Victor Fuentes

02/06/2021

Setup

```
library(data.table)
library(ggplot2)
library(ggpubr)
library(viridis)
library(kknn)
library(boot)
library(MASS)
library(rpart)
library(rpart.plot)
library(kableExtra)
```

1 Loading Data

```
setwd("C:/Users/vfuentesc/OneDrive - The University of Chicago/Winter 2021/Machine Learning/Week 3/HW3-B
train = data.table(read.csv("Bike_train.csv"))
test = data.table(read.csv("Bike_test.csv"))
```

```
train[, log_count := log(count + 1)]
train[, count := NULL]

season_labels = c("Winter", "Spring", "Summer", "Fall")
train[, season := factor(season, levels = 1:4, labels = season_labels)]
train[, weather := factor(weather)]
```

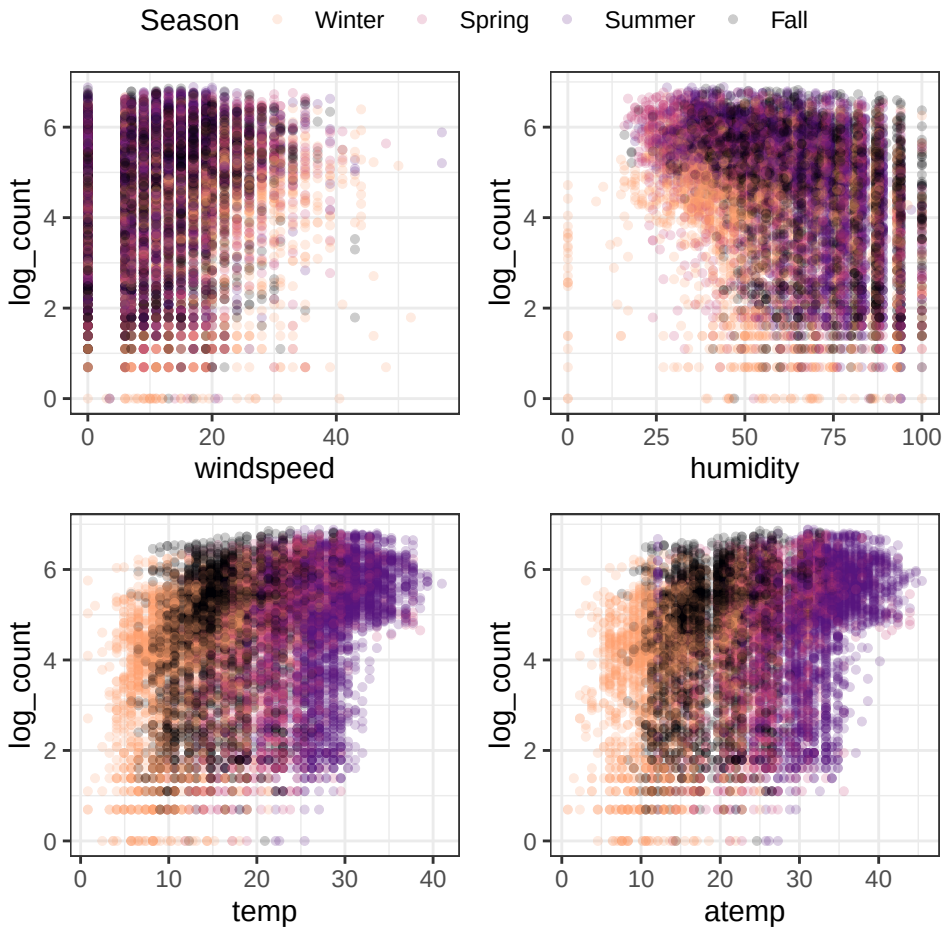
2 Questions

1. Data Exploration

a. Visualize the relationship between count and each one of the following variables on a separate scatter plot: `windspeed`, `humidity`, `temp`, and `atemp`.

```
scatter_plot = function(x_var){
return(ggplot(train, aes(x = get(x_var), y = log_count,color = season, fill = season)) +
         geom_point(shape = 21, size = 1.5, stroke = 0.1, alpha = 0.2) +
         scale_fill_viridis_d(option = "A", end = 0.8, direction = -1) + scale_color_viridis_d(opti
         labs(x = x_var, fill = "Season", color = "Season") + theme_bw())}
```

```
ggarrange(scatter_plot("windspeed"), scatter_plot("humidity"),
          scatter_plot("temp"), scatter_plot("atemp"),
          common.legend = TRUE)
```



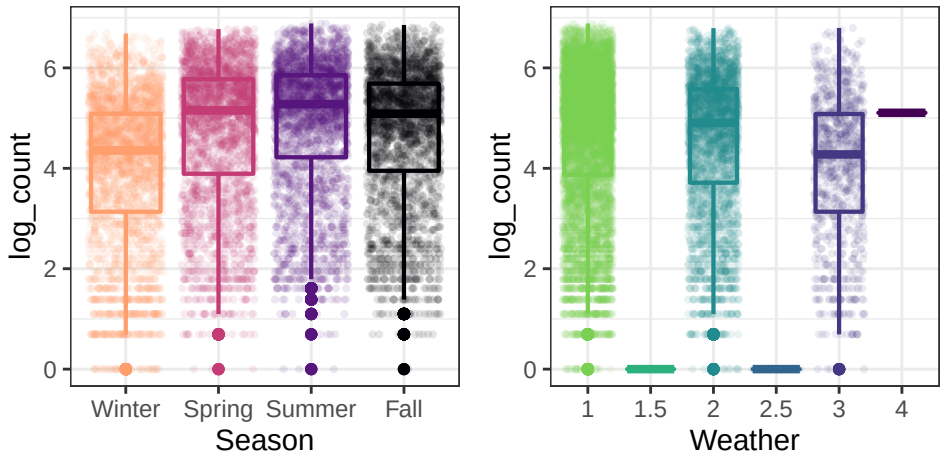
Outliers in ***humidity*** *(atypical zero values) and* ***atemp*** *(very low values atemp for a Summer season).*

b. How does count depend on the season? Consider visualizing this relationship with a boxplot.

```
season_boxplot =
  ggplot(train, aes(x = season, y = log_count)) +
    geom_boxplot(aes(color = season), size = 0.75) +
    geom_jitter(aes(color = season), size = 0.75, alpha = 0.1) +
    scale_color_viridis_d(option = "A", end = 0.8, direction = -1) +
    labs(x = "Season") + theme_bw() + theme(legend.position = "none")

weather_boxplot = ggplot(train, aes(x = weather, y = log_count)) +
    geom_boxplot(aes(color = weather), size = 0.75) +
    geom_jitter(aes(color = weather), size = 0.75, alpha = 0.1) +
    scale_color_viridis_d(option = "D", end = 0.8, direction = -1) +
    labs(x = "Weather") + theme_bw() + theme(legend.position = "none")
```

```
ggarrange(season_boxplot, weather_boxplot)
```



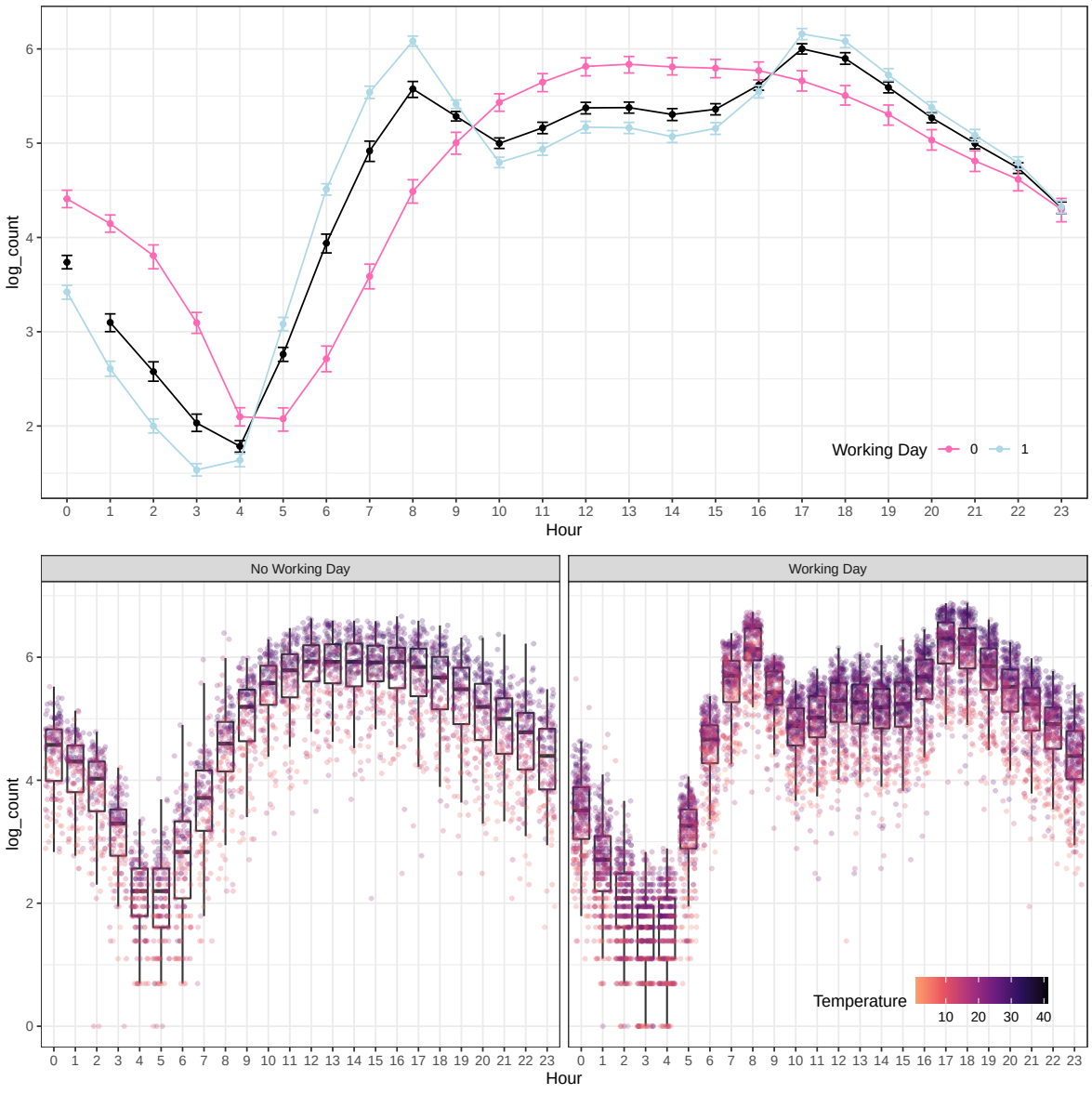
*There is less rentals (**count**) during Winter **season**. On the other hand, **weather** shows irregular values with decimals points.*

c. How does `count` depend on the time of the day (`hour`)? Does this relationship change depending on whether it is a `workingday` or not? A scatterplot could be used to visualize the relationship. You might consider coloring the observations on the scatterplot using the temperature (`temp` or `atemp`) do discern how the temperature affects hourly number of rentals.

```
hour_boxplot =
  ggplot(train, aes(x = factor(hour), y = log_count, group = workingday, color = factor(workingday)
    stat_summary(fun.y = mean, geom = "line", na.rm = TRUE, group = NA, color = "black") +
    stat_summary(fun.y = mean, geom = "point", na.rm = TRUE, group = NA, color = "black") +
    stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width = 0.25, na.rm = TRUE, group = NA
    stat_summary(fun.y = mean, geom = "line") +
    stat_summary(fun.y = mean, geom = "point") +
    stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width = 0.25) +
    scale_color_manual(values = c("hotpink", "lightblue")) +
    labs(x = "Hour", color = "Working Day") + theme_bw() +
    theme(legend.position = c(0.85, 0.10), legend.background = element_rect(fill=NA), legend.direct
```

```
hour_temp_boxplot =
  ggplot(train, aes(x = factor(hour), y = log_count)) +
    geom_boxplot(size = 0.6, color = "gray21", outlier.color = NA) +
    geom_jitter(aes(color = temp), size = 1, alpha = 0.25) +
    scale_color_viridis_c(option = "A", end = 0.8, direction = -1) +
    labs(x = "Hour", color = "Temperature") + theme_bw() +
    facet_grid(~workingday, labeller = labeller(workingday = c("0" = "No Working Day", "1" = "Worki
    theme(legend.position = c(0.85, 0.10), legend.background = element_rect(fill=NA), legend.direct
```

```
ggarrange(hour_boxplot, hour_temp_boxplot, nrow = 2)
```



Bike demand higher during peak hours (7-9am and 4-7pm). But it only applies for working days. During no working days, bike demand is higher from 10am to 6pm. Additionally, the higher the temperature, the higher the demand for bikes irrespective of whether is working day or not.

d. Does the relationship between `count` and `hour` change by season?

e. Does the distribution of hourly number of rentals change between 2011 and 2012? What does this tell you about the rental business?

2. Fitting a *Random Forest* model and a *Boosting* model

a. Create a partial dependence plot for predicted `count` vs each one of the following variables: `windspeed`, `humidity`, `temp`, and `atemp`. Do this for both the *Random Forest* and the *Boosting* model.

b. Build a marginal model that regresses count on each one of the following variables: `windspeed`, `humidity`, `temp`, and `atemp`. You can use whichever nonlinear model you want for these regression tasks. Plot the marginal fits and compare them with partial dependence plots above. How are the plots different and why?

c. Create variable importance plots for the two models in part 2.a. Do the two models rank the variables in the same way?

3. Investigate how predictive each variable is on its own. There are a number of ways to do this. The simplest way would be to regress `count` on each one of the variables separately and evaluate the out-of-sample MSE for each one of the models. For instance, you could fit a regression tree model.

a. What can you say by comparing how predictive each variable is on its own vs the variable importance ranking obtained in the previous question?

b. Why do you think is the reason for the difference?

c. How could you use the variable importance ranking to select variables? We will talk about this in detail in Week 5. Here I want you to think a bit about the variable selection problem and how would you use the tools that you have learnt so far to identify a good set of variables.

4. Build a model to predict the *bikeshare counts* for the hours recorded in the test dataset. Save your predictions to a `.csv` file that you will submit to Kaggle (see Kaggle instruction below.) Provide a write-up that explains how you went about building your model. Attach the code to create the submission `.csv` file as an appendix to your homework submission.

3 Appendix

```
sampleSubmission = data.table(Id = 1:length(yhat),
                              count=yhat)

write.csv(sampleSubmission,
          file = "sampleSubmission.csv",
          row.names = FALSE,
          quote = FALSE)
```