Problem set 2

Matt He

2022-09-11

1.Read in the data and identify the number of rows, features, and the data types.

```
library(readr)
library(tidyverse)
bikes <- read_csv("~/Downloads/bikes_ps.csv")
glimpse(bikes)
```

```
## Rows: 731
## Columns: 10
## $ date        <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-0~
## $ season      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ holiday     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0~
## $ weekday     <dbl> 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4~
## $ weather     <dbl> 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2, 2, 2, 2~
## $ temperature <dbl> 46.71653, 48.35024, 34.21239, 34.52000, 36.80056, 34.88784~
## $ realfeel    <dbl> 46.39865, 45.22419, 25.70131, 28.40009, 30.43728, NA, 28.0~
## $ humidity    <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0.518261~
## $ windspeed   <dbl> 6.679665, 10.347140, 10.337565, 6.673420, 7.780994, 3.7287~
## $ rentals     <dbl> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 12~
```

In this dataset **bikes**, there are 731 rows and 10 columns. So there are 731 observations and 10 variables.

2.Convert data types as appropriate. You may have to make some decisions here. Please write a few sente

```
bikes <- bikes %>%
  mutate_at(vars(season,holiday,weekday,weather), .funs = factor)%>%
  mutate(holiday = ifelse(holiday == 0,'Yes','No'))
```

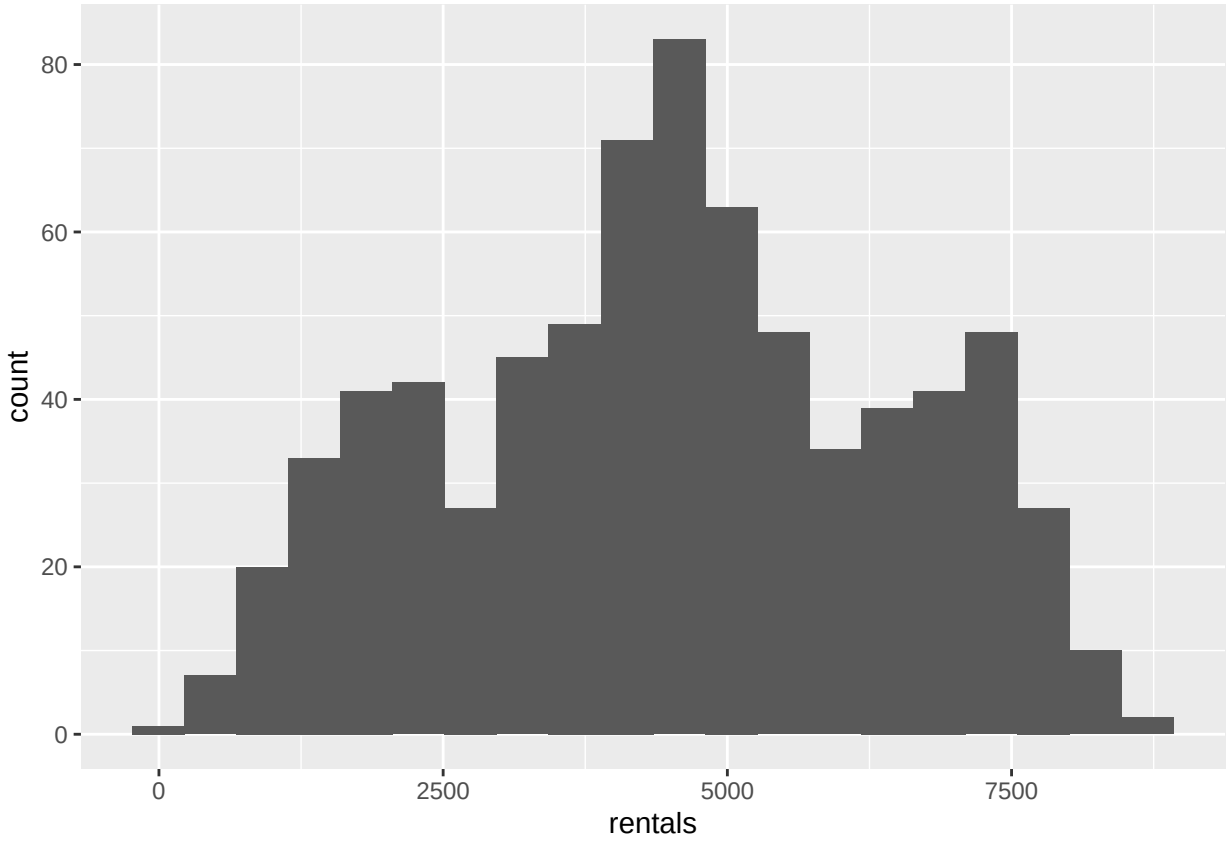
I convert 'season', 'holiday', 'weekday', and 'weather' into factors. For 'holiday', I convert binary 0 and 1 into characters No and Yes. When presenting my data to people who do not read binary, it is easier to understand.

3.Use median imputation to fill in any missing values in continuous variables.

```
bikes <- bikes %>%
  mutate(realfeel = ifelse(is.na(realfeel),median(realfeel,na.rm = TRUE),realfeel))
```

4.Analyze and prepare the target feature: rentals (visualize it, is it skewed? Need to be transformed? A

```
library(ggplot2)
bikes %>%
  ggplot(aes(x = rentals))+
  geom_histogram(bins = 20)
```



The distribution of 'rentals' is not basically bell-shaped. Left and right sides are lower than the middle part. However, there are suddenly sliding down in each side, which makes three peeks in the histogram. I think the 'rentals' needs to be transformed since the range is too large(min is 22 while max is 8714).

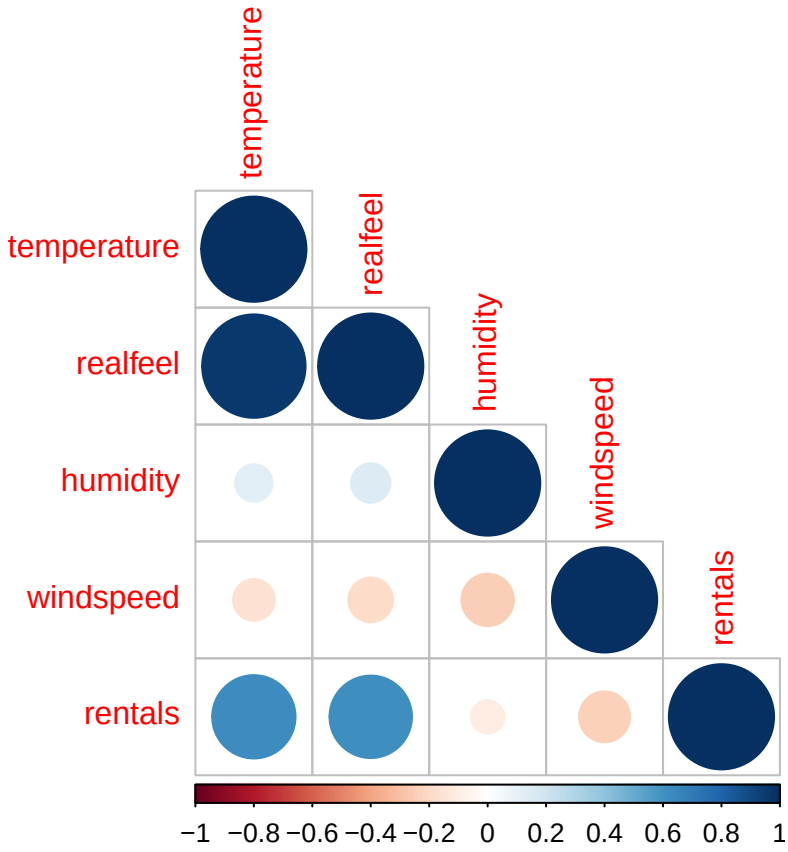
5.Analyze and prepare the non-target features
  1.create a correlation plot of the numeric features, and also a gallery of distribution and correlati

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrMatrix = bikes %>%
  keep(is.numeric) %>%
  cor()

corrplot(corrMatrix,type = 'lower')
```



2.decide if any variables should be dropped because of multicollinearity, etc.

Maybe one of the 'temperature' and 'realfeel' should be dropped, because they are highly correlated to each other and their correlation with 'rentals' are quite similar.

3.Z-score normalize temperature.

```
bikes <- bikes %>%
  mutate(temperature = (temperature-mean(temperature))/sd(temperature))
```

4.Min-max normalize the windspeed variable

```
bikes <- bikes %>%
  mutate(windspeed = ((windspeed-min(windspeed))/(max(windspeed)-min(windspeed)))*(1-0)+0)
```

5.Convert all cateogorical variables into dummy variables (the dummy.data.frame function can make this

```
library(dummy)
```

```
## dummy 0.1.3
```

```
## dummyNews()
```

```
bikes2 <- bikes %>%
  dummy()
```

6.Train a linear regression which uses all the features (except those you might remove) to predict renta

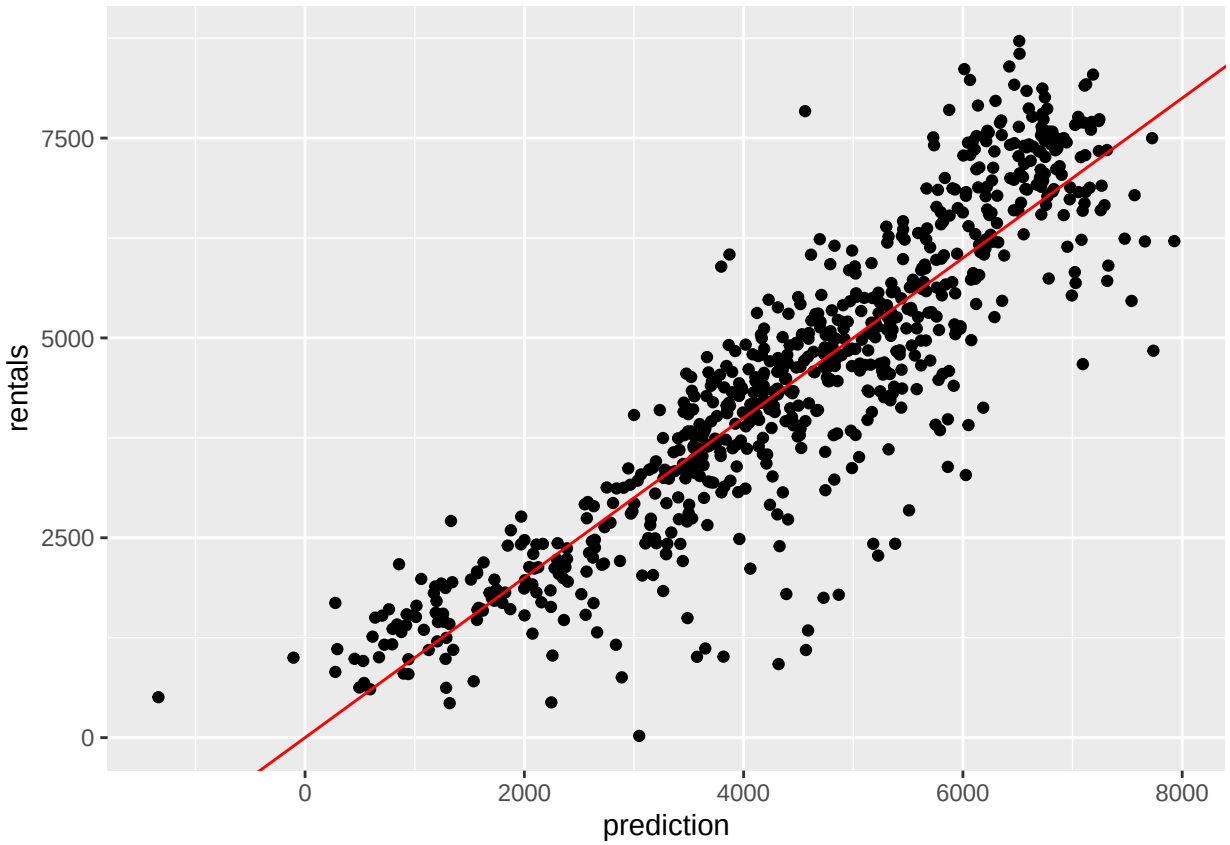
```
linear_model <- lm(rentals ~. , data = select(bikes,-realfeel))
summary(linear_model)
```

```
##
## Call:
## lm(formula = rentals ~ ., data = select(bikes, -realfeel))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3472.6  -386.7    70.3   533.4  3275.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.077e+04  2.666e+03 -26.545  < 2e-16 ***
## date         4.928e+00  1.717e-01  28.700  < 2e-16 ***
## season2      9.004e+02  1.215e+02   7.409 3.60e-13 ***
## season3      1.368e+02  1.607e+02   0.851 0.394812
## season4      4.240e+02  1.101e+02   3.850 0.000129 ***
## holidayYes   7.399e+02  2.039e+02   3.629 0.000305 ***
## weekday1     2.184e+02  1.253e+02   1.743 0.081699 .
## weekday2     3.057e+02  1.224e+02   2.497 0.012737 *
## weekday3     3.638e+02  1.227e+02   2.965 0.003129 **
## weekday4     3.739e+02  1.226e+02   3.049 0.002384 **
## weekday5     4.013e+02  1.227e+02   3.272 0.001119 **
## weekday6     4.184e+02  1.220e+02   3.430 0.000638 ***
## weather2    -4.007e+02  8.729e+01  -4.590 5.24e-06 ***
## weather3    -1.959e+03  2.234e+02  -8.772  < 2e-16 ***
## temperature  9.497e+02  6.023e+01  15.767  < 2e-16 ***
## humidity    -1.648e+03  3.167e+02  -5.203 2.57e-07 ***
## windspeed   -1.377e+03  2.227e+02  -6.183 1.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 881.2 on 714 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7931
## F-statistic: 175.9 on 16 and 714 DF,  p-value: < 2.2e-16
```

```
bikes <- bikes %>%
  mutate(prediction = predict(linear_model,newdata = bikes))
```

7.Visualize the actual rentals and the predicted rentals over time.

```
bikes %>%
  ggplot(aes(y = rentals, x = prediction))+
  geom_point()+
  geom_abline(color = 'red')
```



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

In most parts, the feature I have in the data make sense. However, there are some parts that is not plausible. First, we can see the prediction actually does not start with zero.There should not be negative rentals. Second, there are some another special group that is under the abline. When our model predicts rental should be quite high, but reality is the rental is not as big as we predicted.

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

I believe train a model means to teach the computer to analyse the certain way how something work. For example, in this **bikes** dataset, we give the machine different information about the weather, weekday and so on. Therefore, the computer, according to the relationship between target feature and these information, can predict when we have similar information again. Try to imagine we made a good rentals when it is a sunny Monday, the next sunny Monday comes, we can predict confidently, the rental for that Monday must be similar to the last Monday.

10.Discuss which preparations you did were required to make the learning algorithm work, and which were

Transforming the data is not strictly required, but it is a really good idea. If we do not transform the data with too big of ranges, we might be confused with the outcome and the model we create.