Linear Regression Model Project

Jason Barnes

5/21/2019

Linear Regression Model

(Bikeshare Challenge)

You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.1
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts --------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(knitr)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
#devtools::install_github('yihui/tinytex')
library(tinytex)
library(rmarkdown)
```

Data

Data can be downloaded at https://www.kaggle.com/c/bike-sharing-demand/data#

```
bike <-  read.csv("Project/bikeshare.csv")
```

Explore the data

```
headbike <- head(bike)
kable(headbike) %>%
    kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	regist
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0000	3	
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0000	8	
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0000	5	
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0000	3	
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0000	0	
2011-01-01 05:00:00	1	0	0	2	9.84	12.880	75	6.0032	0	

Check the structure of bike

```
str(bike)
```

```
## 'data.frame':    10886 obs. of  12 variables:
##  $ datetime  : Factor w/ 10886 levels "2011-01-01 00:00:00",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
##  $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
##  $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
##  $ windspeed : num  0 0 0 0 0 ...
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
```

Factors

Since weather and season are factors we will want to change the data to as.factor

```
bike$season <- as.factor(bike$season)
bike$weather <- as.factor(bike$weather)
```

Now we will call structure aagin so we can verify the changes

```
str(bike)
```

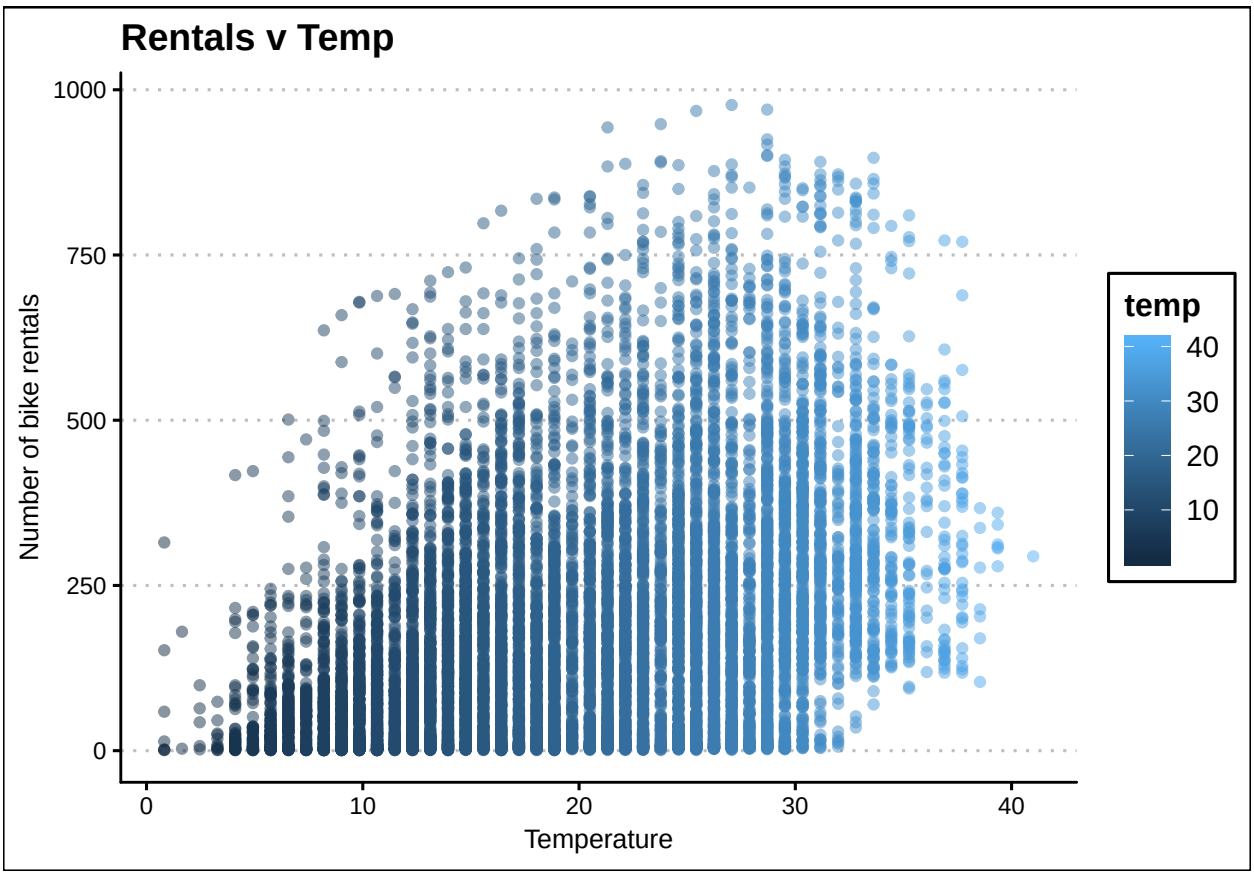
```
## 'data.frame':    10886 obs. of  12 variables:
##  $ datetime  : Factor w/ 10886 levels "2011-01-01 00:00:00",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ season    : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weather   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
##  $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
##  $ humidity  : int  81 80 80 75 75 75 80 86 75 76 ...
##  $ windspeed : num  0 0 0 0 0 ...
```

```
##  $ casual    : int  3 8 5 3 0 0 2 1 1 8 ...
##  $ registered: int  13 32 27 10 1 1 0 2 7 6 ...
##  $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
```

Visualize the data

count vs temp

```
ggplot(bike, aes(temp, count)) +
    geom_point(aes(color = temp), alpha = .5) +
    labs(x = "Temperature", y = "Number of bike rentals", title = "Rentals v Temp") +
    theme_clean()
```

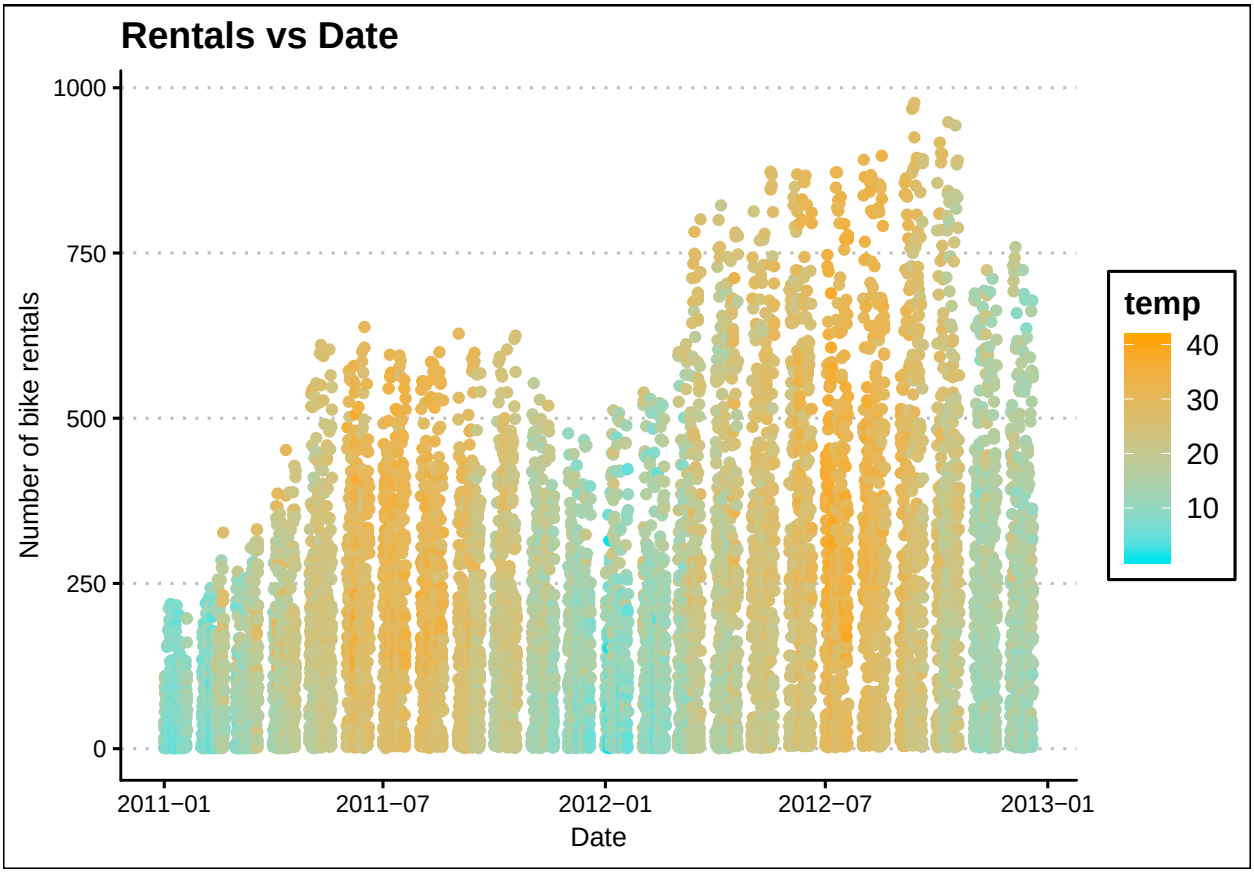


In order to vizualize Need to turn datetime into POSIXct data field

```
bike$datetime <- as.POSIXct(bike$datetime)
```

count vs datetime

```
ggplot(bike, aes(datetime, count)) +
    geom_point(aes(color = temp)) +
    scale_color_continuous(low = "turquoise2", high = "orange") +
    labs(x = "Date", y = "Number of bike rentals", title = "Rentals vs Date") +
    theme_clean()
```



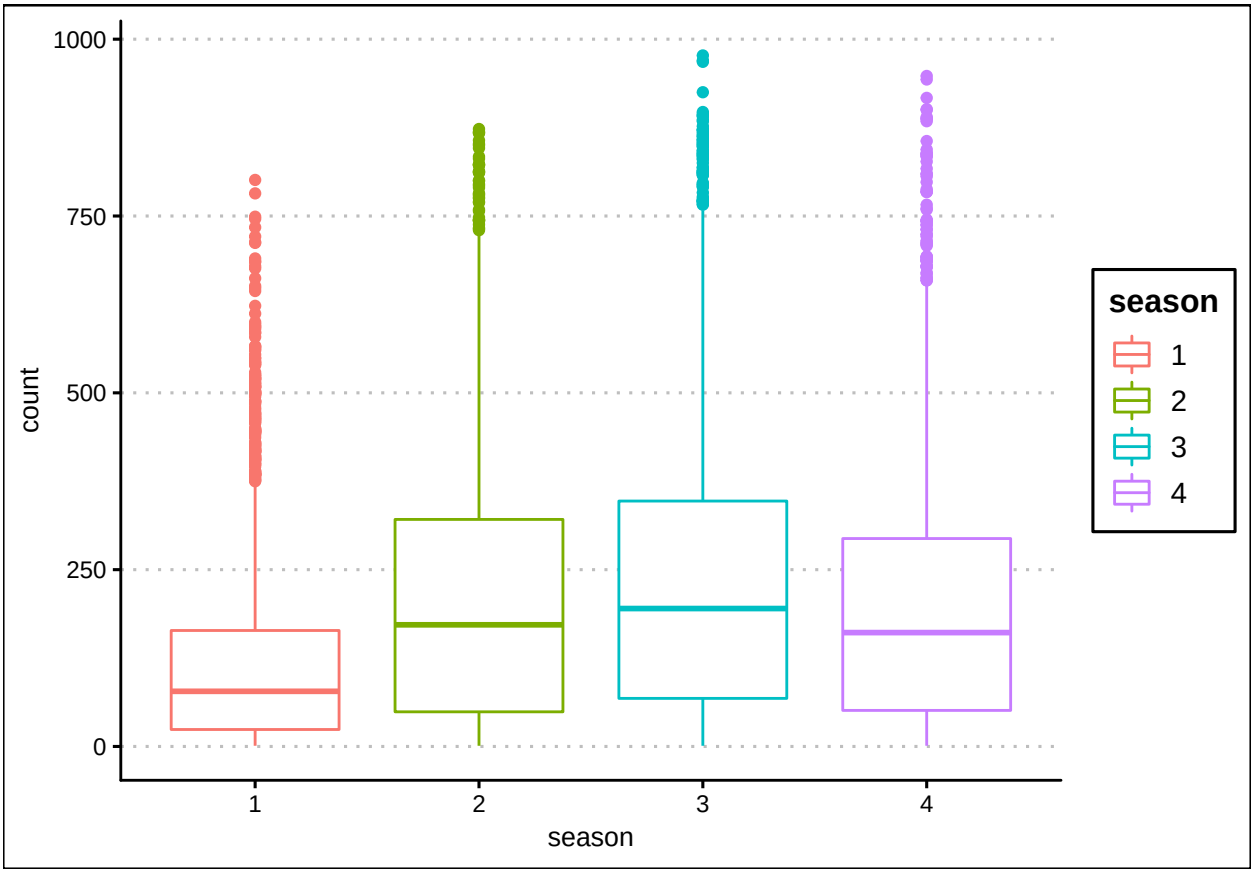
Coorelation between number of bikes rented and the temperature

```
cor(bike[,c("temp","count")])
```

```
##             temp     count
## temp  1.0000000 0.3944536
## count 0.3944536 1.0000000
```

Expolore seasonality with a boxplot

```
ggplot(bike, aes(season, count)) +
    geom_boxplot(aes(color = season)) +
    theme_clean()
```



Add an hour column

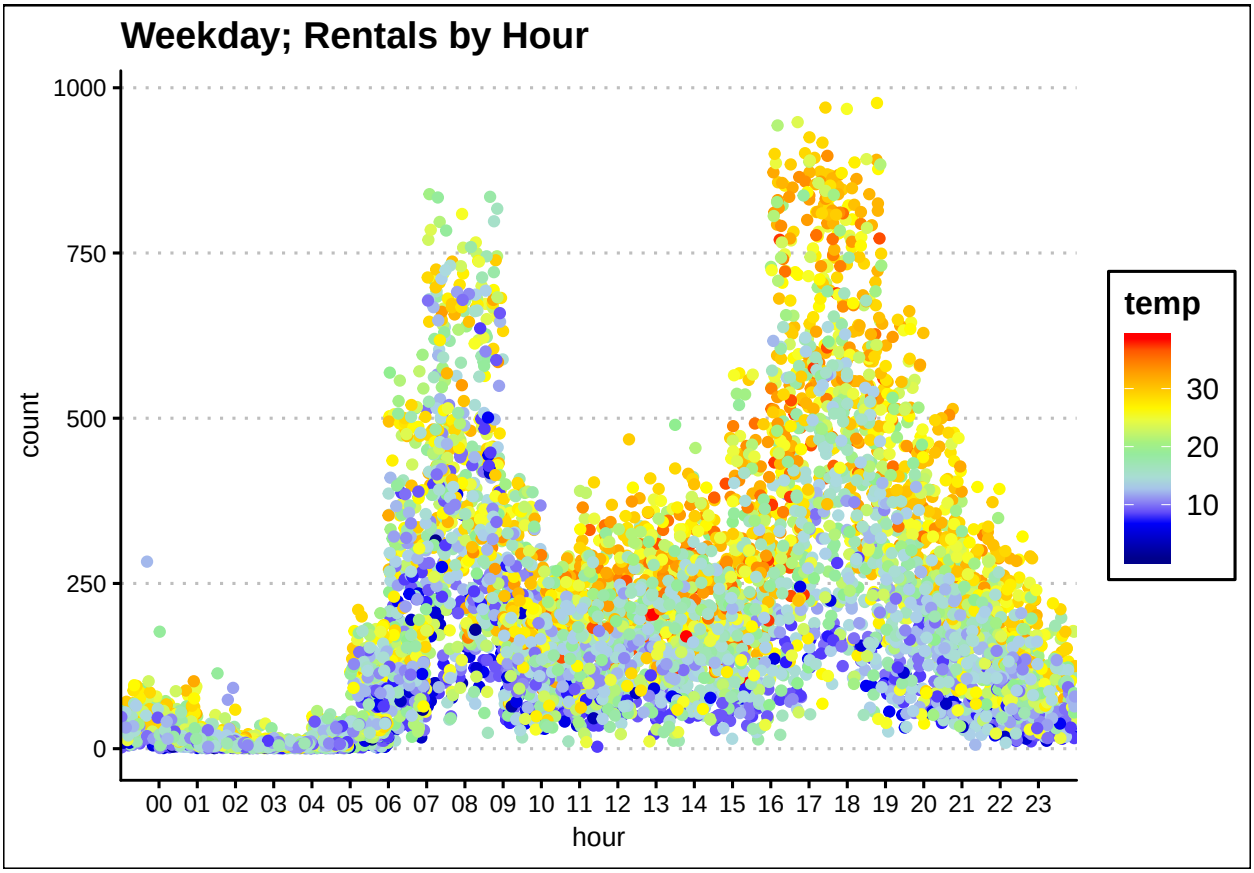
```
bike$hour <- sapply(bike$datetime, function(x){format(x,"%H")})
```

Segment data into weekday and weekend

```
workday <- bike %>%
    filter(workingday == 1)
nonworkday <- bike %>%
    filter(workingday == 0)
```

Visulaize weekday

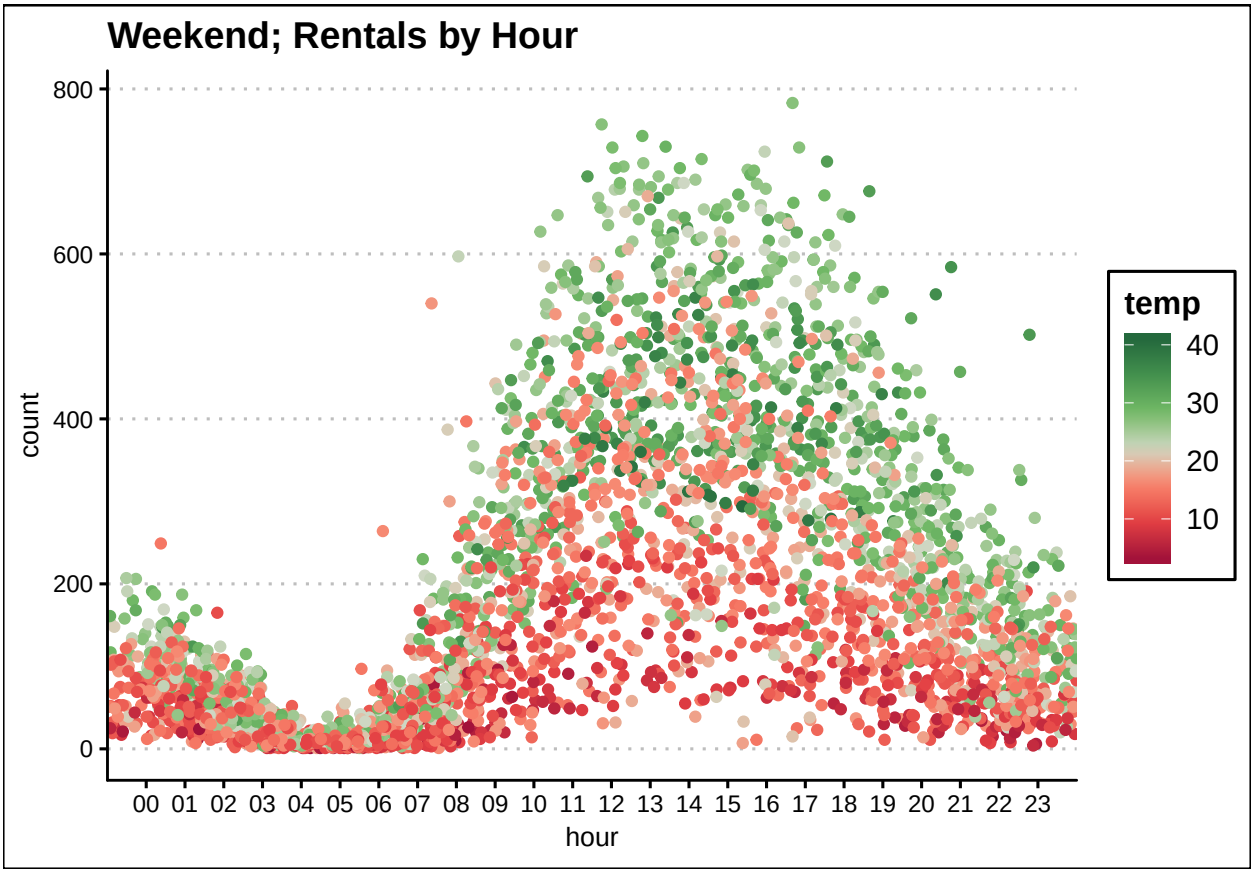
```
ggplot(workday, aes(hour, count)) +
    geom_point(aes(color = temp), position = position_jitter(width = 1, height = 0)) +
    scale_color_gradientn(colours = c('dark blue','blue','light blue','light green','yellow','orange','
    labs(title = "Weekday; Rentals by Hour") +
    theme_clean()
```



Notice the peak in the morning before work and the evening after work?

Visualize the weekend

```
ggplot(nonworkday, aes(hour, count)) +
    geom_point(aes(color = temp), position = position_jitter(width = 1, height = 0)) +
    scale_color_gradient2_tableau(palette = "Red-Green Diverging") +
    labs(title = "Weekend; Rentals by Hour") +
    theme_clean()
```



Building the Model with just temperature as an attribute

```
temp.model <- lm(formula = count ~ temp, data = bike )
summary(temp.model)
```

```
##
## Call:
## lm(formula = count ~ temp, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.32 -112.36  -33.36   78.98  741.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.0462     4.4394   1.362    0.173
## temp          9.1705     0.2048  44.783   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 166.5 on 10884 degrees of freedom
## Multiple R-squared:  0.1556, Adjusted R-squared:  0.1555
## F-statistic:  2006 on 1 and 10884 DF,  p-value: < 2.2e-16
```

Interpeting the Model

Interpreting the intercept: It is the value of y when x=0.Thus, it is the estimated number of

rentals when the temperature is 0 degrees Celsius. Note: It does not always make sense to interpret the intercept.

Interpreting the "temp" coefficient: It is the change in y divided by change in x, or the "slope". Thus, a temperature increase of 1 degree Celsius is associated with a rental increase of 9.17 bikes. This is not a statement of causation. Coefficient would be negative if an increase in temperature was associated with a decrease in rentals.

Predict the number of rentals if the temperature was 25 degrees Celsius

```
# y = mx + b
9.17 * (25) + 6.0462
```

```
## [1] 235.2962
```

Building the final model

- Attributes
  - season
  - holiday
  - workingday
  - weather
  - temp
  - humidity
  - windspeed
  - hour (factor)

Convert hour to a numeric value

```
bike$hour <- sapply(bike$hour, as.numeric)
```

Model

```
model_bike <- lm(count ~ . -casual - registered - datetime - atemp, bike)
summary(model_bike)
```

```
##
## Call:
## lm(formula = count ~ . - casual - registered - datetime - atemp,
##     data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -342.22  -96.15  -31.35   54.23  673.28
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.05365    8.86641   3.390 0.000702 ***
## season2      15.34914    5.15784   2.976 0.002928 **
## season3     -12.80216    6.56666  -1.950 0.051253 .
## season4      66.85422    4.31557  15.491  < 2e-16 ***
## holiday      -9.50871    8.71884  -1.091 0.275476
## workingday   -1.49264    3.11903  -0.479 0.632262
## weather2     10.12225    3.42440   2.956 0.003124 **
## weather3    -29.62495    5.77708  -5.128 2.98e-07 ***
## weather4    104.05163  146.62722   0.710 0.477946
## temp          9.16839    0.30887  29.684  < 2e-16 ***
## humidity     -2.07404    0.09116 -22.752  < 2e-16 ***
```

```
## windspeed     0.18471    0.18526   0.997 0.318769
## hour          7.41472    0.21733  34.117  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.5 on 10873 degrees of freedom
## Multiple R-squared:  0.3463, Adjusted R-squared:  0.3456
## F-statistic:   480 on 12 and 10873 DF,  p-value: < 2.2e-16
```