Bike Rental Report

Louise Braithwaite

31/07/2020

1. Business Overview

Over a two year period a bike rental company has collected data on the number of daily rentals, daily weather conditions (temperature, humidity, wind speed and precipitation type) and day type (season, year, month, holiday, weekday, working day). The company has commissioned the NICD to analyse and deliver business insights that will help them understand how the change in weather effects the number of bikes they rent each day.

1.1 Data mining goals and success criteria

This project will focus on data description and summarisation of the data set to help develop insights into the impact weather has on bike rentals.

The goal of this project is to:

Identify how the change in weather effects the number of bikes rented each day.

1.2 Requirements, assumptions, and constraints

Requirements:

- A report with a maximum of 500 words and 2 figures, detailing how the data has been interpreted and what insights have been gained.
- A Git repository of the work completed
- A ProjectTemplate folder containing reports and source code

Assumptions and constraints:

- We can only work with the data set provided
- No meta date has been provided. All data assumptions should be logged so they can be verified later

2. Data Overview and Preparation

A data set was provided by the bike rental company as a csv file. The data set is complete with no missing values.

The next step was to investigate what information is contained within the data set and evaluate which variables relate to the project goal. The section below contain:

- a snapshot of the first 5 observations of the data file. This provides the column headers and vector class
- a summary table of the variables in the provided data set
- a list of the key variables associated with the project goal
- a summary of data assumptions
- data preparation notes
- table summarising the final two data frames
- a summary table of the variables in the final data set

2.1 Data Overview

Review the variables in the original bike.rental.data.csv file. In R the original data (bike.rental.data.csv) is saved as a data frame named 'rental.data'.

```
head(rental.data, 5)
```

```
## # A tibble: 5 x 12
##   season    yr mnth  holiday weekday workingday weathersit  temp   hum windspeed
##   <chr>  <dbl> <chr> <chr>   <chr>   <chr>      <chr>      <dbl> <dbl>     <dbl>
## 1 SPRING  2011 JAN   NO HOL~ SAT     NO WORKIN~ MISTY       8.18  80.6      10.7
## 2 SPRING  2011 JAN   NO HOL~ SUN     NO WORKIN~ MISTY       9.08  69.6      16.7
## 3 SPRING  2011 JAN   NO HOL~ MON     WORKING D~ GOOD        1.23  43.7      16.6
## 4 SPRING  2011 JAN   NO HOL~ TUE     WORKING D~ GOOD        1.4   59.0      10.7
## 5 SPRING  2011 JAN   NO HOL~ WED     WORKING D~ GOOD        2.67  43.7      12.5
## # ... with 2 more variables: cnt <dbl>, days_since_2011 <dbl>
```

.

Table 1: Summary of the variables from bike_rental_data csv file

Column Header	Class	Example	Description
season	character	"SPRING"	The season, spring, summer, fall or winter
yr	integer	2011	The year, either 2011 or 2012
mnth	character	"JAN"	The month, all in shortened character form (e.g. "JAN", "FEB", "MAR")
holiday	character	"NO HOLIDAY"	Whether the day is a holiday "HOLIDAY" or not "NO HOLIDAY"
weekday	character	"SAT"	The day of the week, all in shortened character form (e.g. "SAT", "SUN", "MON")
workingday	character	"NO WORKING DAY"	Whether it is a "WORKING DAY" or a "NO WORKING DAY"
weathersit	character	"MISTY"	Weather category ("GOOD", "MISTY" or "RAIN/SNOW/STORM")
temp	numeric	8.18	Temperature (degrees celsius, °C)
hum	numeric	80.6	Relative humidity (%)
windspeed	numeric	10.7	Wind speed (mph)
cnt	integer	985	The number of bikes rented that day (rental count)
days_since_2011	integer	0,1,2	A sequence of numbers starting a 0 on 1 January 2011 and increasing by 1 each day

2.1.1 Key Variables There are five variables we can use to help us determine how weather can effect the number of bicycle rentals. They are:

- cnt
- weathersit
- temp
- hum
- windspeed

2.1.1 Data Assumptions The following data assumptions were made. They should be verified with the data owner at the next opportunity.

- Temperature is measured in degrees celsius (°C)
- Humidity measure is relative humidity (%)
- Wind speed is measured in miles per hour (mph)
- Temperature, humidity and wind speed can be rounded to integers, as this level of detail is most appropriate and will make analysis and understanding easier.
- Weathersit variables are an indication of precipitation, "GOOD" can also be referred to as "NO.RAIN".

2.2 Data Preparation Notes

The following data preparation tasks were performed to ease the exploratory data analysis process.

1. To ease navigate through the data, create proper date variables to replace 'yr', 'month', 'weekday' and 'days_since_2011'.

i) Create date column, use the 'days_since_2011' variable to quality check the new date variable. Eg. observation 31 of 'days_since_2011' is 30 but this equates to 31 January 2011
ii) Introduce the lubridate package to create new date variables, 'just.day', 'just.dayofweek', 'just.dayofweek2' (character labels), 'just.month', 'just.month2' (character labels) and 'year'
iii) Remove the original date columns, which are no longer needed ('yr', 'mnth', 'weekday')

2. Convert categorical character variables to factors ('season', 'holiday', 'workingday' and 'weathersit')
3. Rename variables for clarity and to create naming convention ('cnt' to 'rental.count', 'weathersit' to 'weather.category', 'temp' to 'temperature, 'hum' to 'humidity', 'windspeed' to 'wind.speed', 'workingday' to 'working.day' and 'days_since_2011' to 'days.since.2011')
4. Round the 'temperature', 'wind.speed' and 'humidity' columns so they become integers
5. Create new column called 'precipitation' using the data from 'weather.categories' and change the characters to numbers to reflect the factor levels
6. Reorder variables so the response variable is listed first

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

2.2.1 Updated Data Frame .

Table 2: Summary of the variables from data frame after data has been prepared

Column Header	Class	Example	Description
rental.count	integer	985	The number of bikes rented that day
precipitation	numeric	1	Weather category (1 = "GOOD", 2 = "MISTY" and 3 = "RAIN/SNOW/STORM")
temperature	numeric	8.18	Temperature (degrees celsius, °C)
humidity	numeric	80.6	Relative humidity (%)
wind.speed	numeric	10.7	Wind speed (mph)
weather.category	character	"MISTY"	Weather category ("GOOD", "MISTY" or "RAIN/SNOW/STORM")
season	character	"SPRING"	The season, spring, summer, fall or winter
holiday	character	"NO HOLIDAY"	Whether the day is a holiday "HOLIDAY" or not "NO HOLIDAY"
working.day	character	"NO WORKING DAY"	Whether it is a "WORKING DAY" or a "NO WORKING DAY"
days.since.2011	integer	0,1,2	A sequence of numbers starting a 0 on 1 January 2011 and increasing by 1 each day
date	date	"2011-01-01"	The date
just.day	integer	1	The day of the month
just.dayofweek	numeric	1	The day, 1-7 (1 = Sunday, 2 = Monday, etc)
just.dayofweek2	factor	"Sun"	The day, all in shortened character form (e.g. "Sun", "Mon", "Tue")
just.month	numeric	"JAN"	The month, 1-12 (1 = Jan, 2 = Feb etc)
just.month2	factor	"Jan"	The month, all in shortened character form (e.g. "Jan", "Feb", "Mar")
year	numeric	2011	The year, either 2011 or 2012

.
.
.

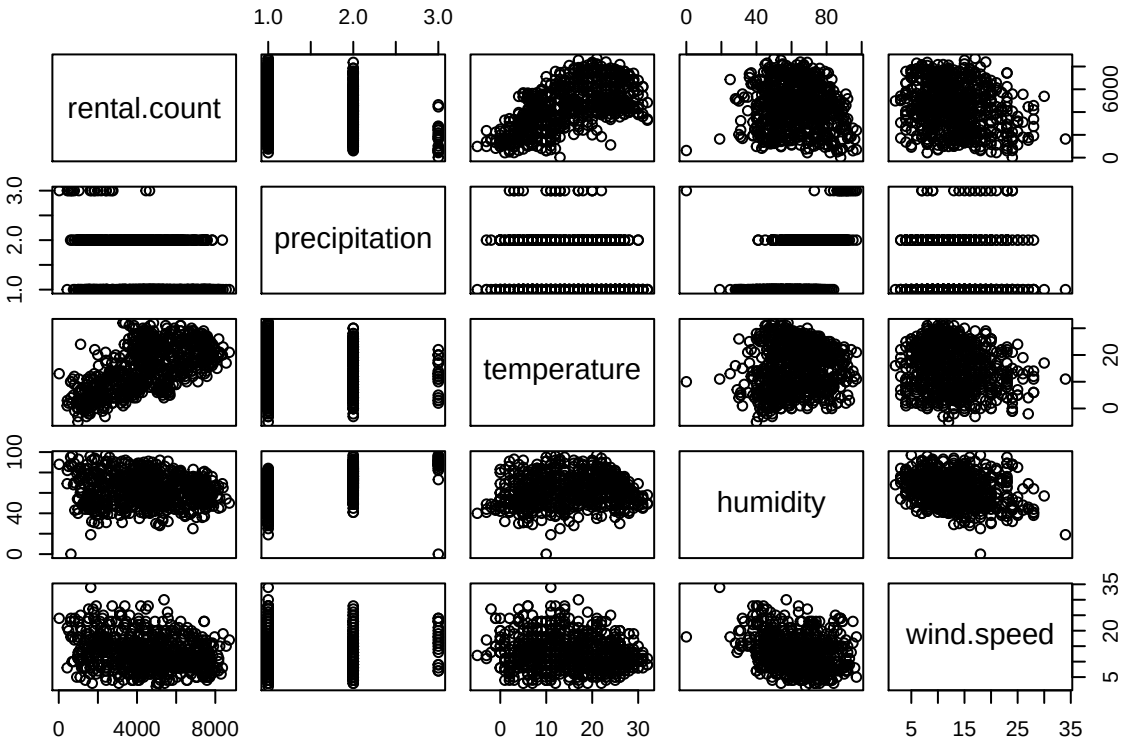
2.2.2 Final Data Frames After preparing the data we are left with two data frames.

Table 3: Summary of the two final data frames

Data Frame	Description	Dimensions
rental.data	The original data	731 obs. and 12 variables
bike.rental.data	Data preparation steps 1-6 applied	731 obs. and 17 variables

3. Exploratory Data Analysis

Pairs Plot Create a pairs plot of the variables relating to rental count and weather.



Findings from pairs plot:

- There appears to be a moderately strong and positive linear correlation between rental.count and temperature; * as temperature increases so does rental.count.
- There appears to be some indication of a negative linear correlation between rental.count and humidity; as humidity increases rental.count decreases.
- There appears to be some indication of a negative linear correlation between rental.count and wind speed; as wind speed increases rental.count decreases.
- The pairs plot indicates that the weather.category is an indicator of precipiation, as the "MISTY" and "RAIN/SNOW/STORM" categories have higher recording of humidity. *It seems safe to assume that good indicates no rain.*

```
# Review the categorical precipitation data
table(bike.rental.key$precipitation)
```

```
##
##   1   2   3
## 463 247  21
```

From the summary above table we can see that weather has been categorised as "GOOD" 63% of the time. 34% of the time the weather is categorised as "MISTY" and 3% of the time the weather is categorised as "RAIN/SNOW/STORM".

Correlation Matrix Use the correlation matrix to quantify the strength of a linear relationship between pairs of variables.

```
# calculate the correlation matrix
cor(bike.rental.key)
```

```
##                rental.count precipitation temperature   humidity  wind.speed
## rental.count     1.0000000   -0.29739124   0.6277925 -0.1015581 -0.23714189
## precipitation   -0.2973912    1.00000000  -0.1210229  0.5918735  0.03884577
## temperature      0.6277925   -0.12102285   1.0000000  0.1266572 -0.16289184
## humidity        -0.1015581    0.59187347   0.1266572  1.0000000 -0.24768354
## wind.speed      -0.2371419    0.03884577  -0.1628918 -0.2476835  1.00000000
```

The correlation matrix confirms some of our previous observations:

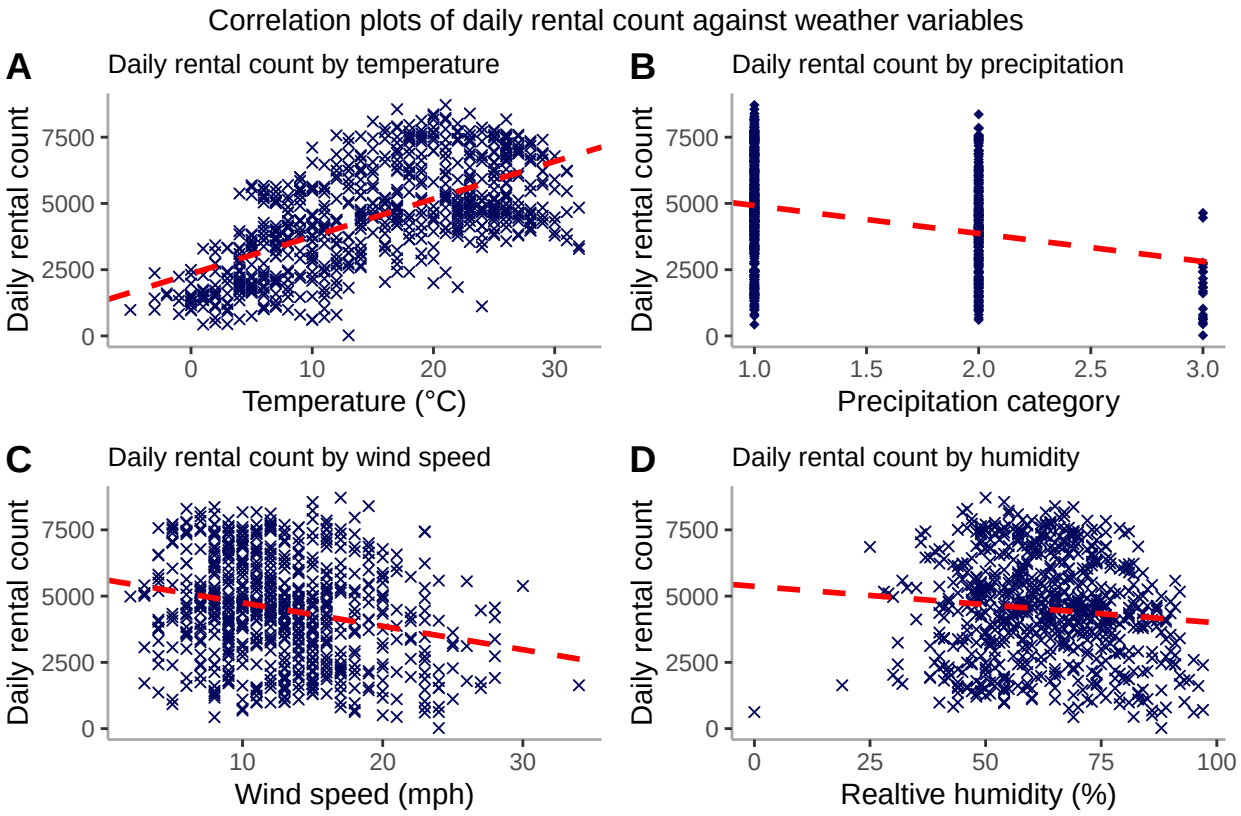
- The correlation between rental.count and temperature, 0.63, indicates a moderately strong positive linear relationship
- The correlation between rental.count and weather.catergory, -0.3, suggests there is some indication of a negative linear relationship between the two variables. This supports the assumption that as the weather goes from no rain ("GOOD"), to "MISTY", to "RAIN/SNOW/STORM" the rental.count will decrease.
- The correlation between weather.category and humidity, 0.59, indicates a moderately strong positive linear relationship. Observation makes sense, as we would expect the air to be more humid if the weather is misty or it is precipitating.

- The correlation between rental.count and wind.speed, -0.24, and rental.count and humidity, -0.1, are negative but are too low to indicate a linear relationship

Correlation plots The correlation plots are ordered to reflect the weather variables' degree of correlation. Temperature has the greatest linear correlation to rental count, then precipitation, wind speed and finally humidity.

```
## Warning: package 'cowplot' was built under R version 4.0.2
```

```
## Warning: package 'ggpubr' was built under R version 4.0.2
```

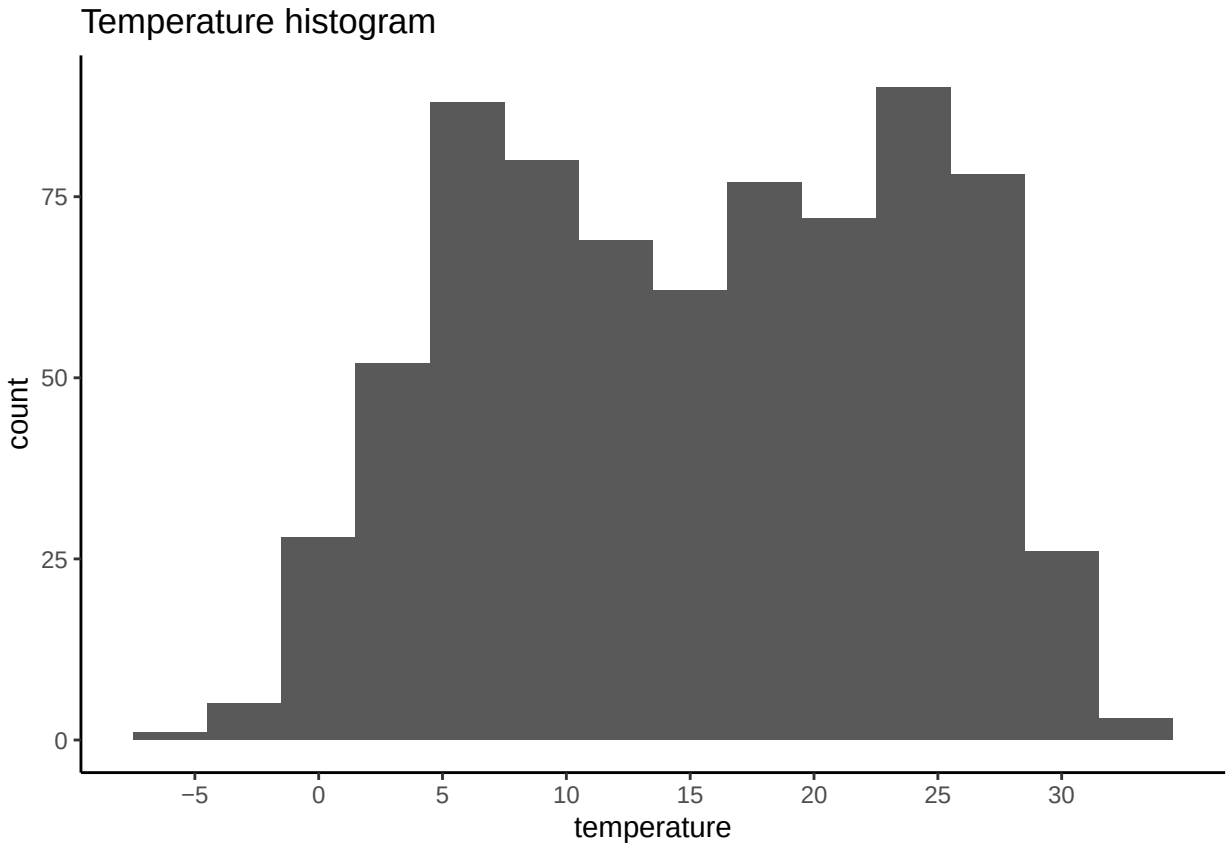


Red dashed line = least square regression line

Summary

The key finding from this analysis is that temperature has the most significant linear relationship with rental count. We will explore this relationship further in the next section.

Temperature The mean temperature across the two years is 15 degrees, however the histogram shows that the distribution of the temperature data is bimodal. The first peak is at 12 degrees and the other at around 31 degrees.

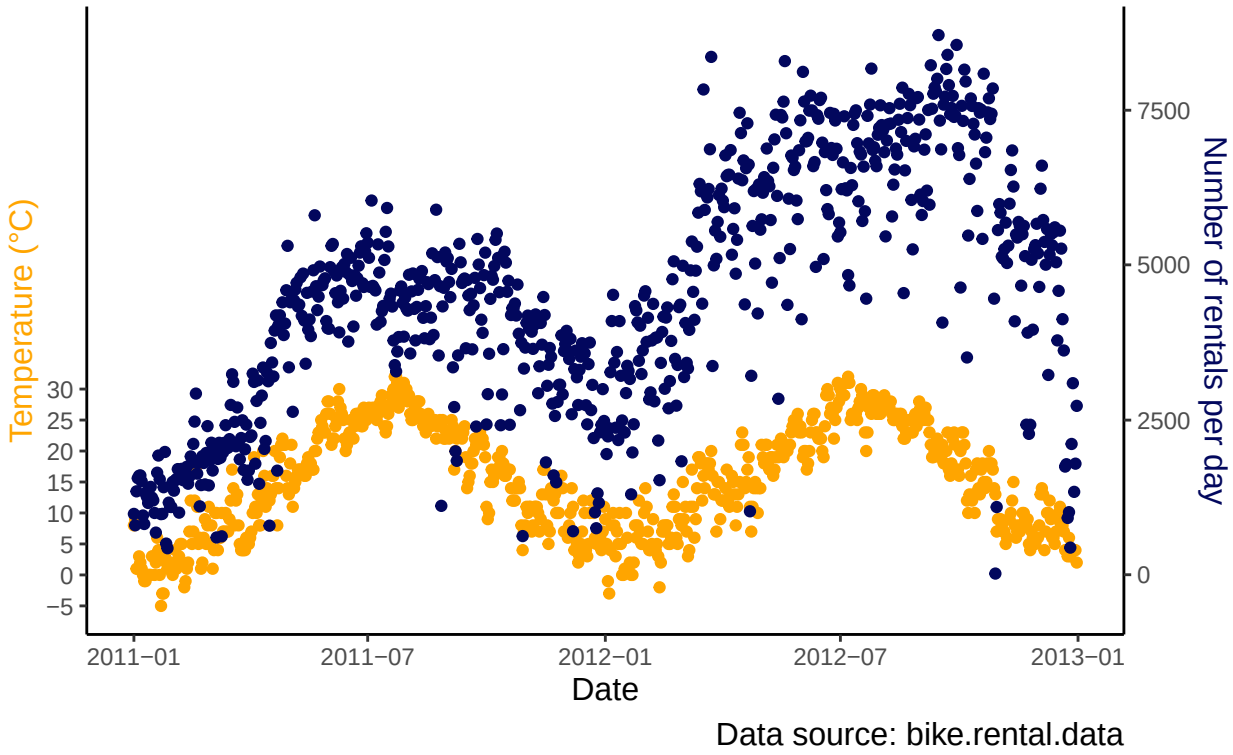



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -5.00    8.00   15.00   15.28   23.00   32.00
```

Let's plot temperature and rental count across the year onto one plot (figure 1).

Figure 1

Rental count and temperature per day



The daily rental count follows the temperature pattern across the two years.

Figure 2

Median number of rentals per day by temperature and year

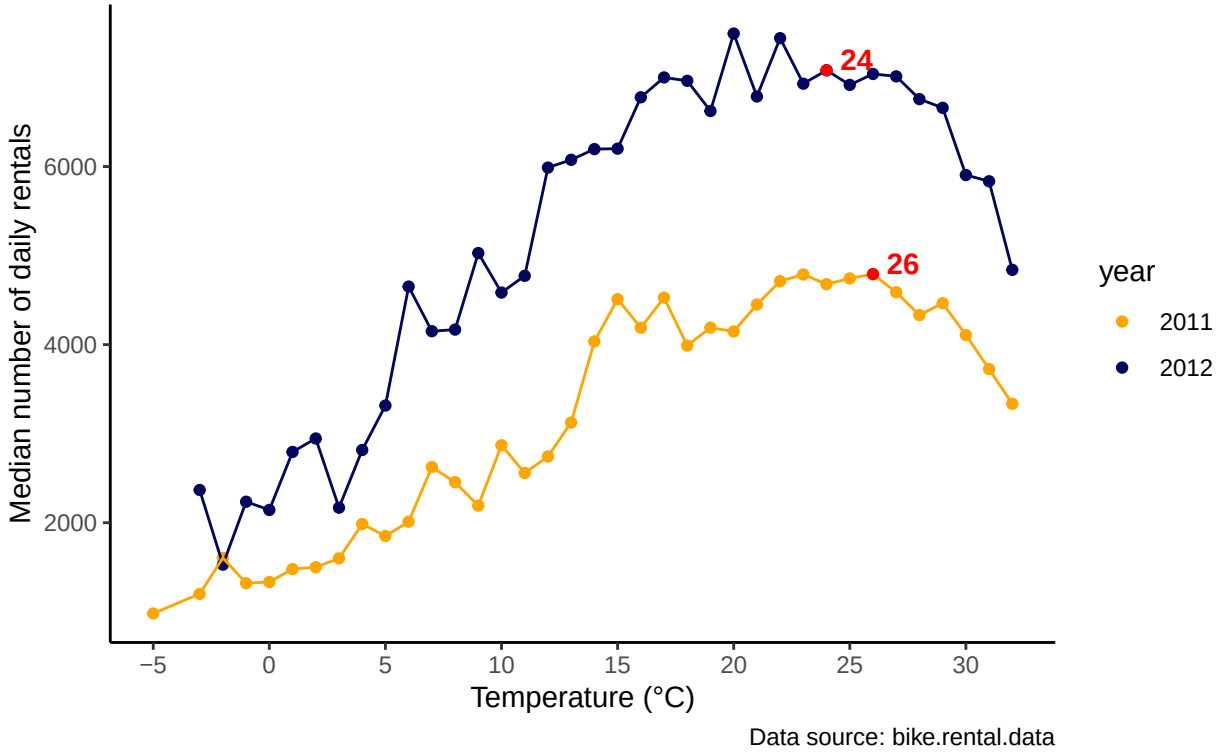


Figure 2 shows that the average number of daily rentals increases with the tempertaure until around 24-26 degrees where it starts decreasing.