Data cleaning bike data

Data Cleaning

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

Information about the data

The data include daily bike rental counts (by members and casual users) of Capital Bikeshare in Washington, DC in 2011 and 2012 as well as weather information on these days. The data is contained in the `dsbox` package and is called `dcbikeshare`.

The original data sources are http://capitalbikeshare.com/system-data and http://www.freemeteo.com.

The codebook is below:

Variable name	Description
`instant`	record index
`dteday`	date
`season`	season (1:winter, 2:spring, 3:summer, 4:fall)
`yr`	year (0: 2011, 1:2012)

Variable name	Description
`mnth`	month (1 to 12)
`holiday`	whether day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
`weekday`	day of the week
`workingday`	if day is neither weekend nor holiday is 1, otherwise is 0.
`weathersit`	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
`temp`	Normalized temperature in Celsius. The values are divided by 41 (max)
`atemp`	Normalized feeling temperature in Celsius. The values are divided by 50 (max)
`hum`	Normalized humidity. The values are divided by 100 (max)
`windspeed`	Normalized wind speed. The values are divided by 67 (max)
`casual`	Count of casual users
`registered`	Count of registered users
`cnt`	Count of total rental bikes including both casual and registered

Load data

```r
bike<-read.csv("~/Downloads/bike+sharing+dataset/day.csv")
```

Mutate categorical variables to factors with informative levels from integers

Season, yr, mnth, holiday, weekday, workningday and wethersit are all categorical variables. In the original dataset these are treated as numerical integers. We need to change them so that R understands that they are categorical. To fin which weekday corresponds to which I searched “calender 2011-01-01” on google, it was a Saturday.

```r
bike <- bike %>%
  mutate(
    weekday=case_when(
      weekday==6~"sat",
      weekday==0~"sun",
      weekday==1~"mon",
      weekday==2~"tue",
      weekday==3~"wed",
      weekday==4~"thu",
      weekday==5~"fri",
    ),
    mnth = case_when(
      mnth==1 ~"jan",
      mnth==2 ~ "feb",
      mnth==3 ~ "mar",
      mnth==4 ~ "apr",
      mnth==5 ~ "may",
      mnth==6 ~"jun",
      mnth==7 ~"jul",
      mnth==8 ~"aug",
      mnth==9 ~"sep",
      mnth==10 ~"okt",
      mnth==11 ~"nov",
      mnth==12 ~"dec"
    ),
    season = case_when(
      season == 1 ~ "winter",
      season == 2 ~ "spring",
      season == 3 ~ "summer",
      season == 4 ~ "fall"
    ),
    holiday = ifelse(holiday == 0, "no", "yes"),
    workingday = ifelse(workingday == 0, "no", "yes"),
    yr = ifelse(yr == 0, "2011", "2012"),
    weathersit = case_when(
      weathersit == 1 ~ "clear",
      weathersit == 2 ~ "mist",
      weathersit == 3 ~ "light precipitation",
      weathersit == 4 ~ "heavy precipitation"
    ))

glimpse(bike)
```

```
Rows: 731
Columns: 16
$ instant    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ dteday     <chr> "2011-01-01", "2011-01-02", "2011-01-03", "2011-01-04", "20~
$ season     <chr> "winter", "winter", "winter", "winter", "winter", "winter",~
$ yr         <chr> "2011", "2011", "2011", "2011", "2011", "2011", "2011", "20~
$ mnth       <chr> "jan", "jan", "jan", "jan", "jan", "jan", "jan", "jan", "ja~
$ holiday    <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no",~
$ weekday    <chr> "sat", "sun", "mon", "tue", "wed", "thu", "fri", "sat", "su~
$ workingday <chr> "no", "no", "yes", "yes", "yes", "yes", "yes", "no", "no", ~
$ weathersit <chr> "mist", "mist", "clear", "clear", "clear", "clear", "mist",~
$ temp       <dbl> 0.3441670, 0.3634780, 0.1963640, 0.2000000, 0.2269570, 0.20~
$ atemp      <dbl> 0.3636250, 0.3537390, 0.1894050, 0.2121220, 0.2292700, 0.23~
$ hum        <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0.518261,~
$ windspeed  <dbl> 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.1869000, 0.08~
$ casual     <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25, 38, 54~
$ registered <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1280, 122~
$ cnt        <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 126~
```

We can see that the categorical variables now are of type characters, they should be factors, this can also be changed with mutate. If we want to also decide which should be the reference level by using fct_relevel() to do both at the same time. The first level will be the reference level and then they will appear in order after that in the regression model. I will use as.Date to convert the dteday variable into something that R recognizes as a date.

```r
bike <- bike %>%
  mutate(
    dteday=as.Date(dteday),
    mnth=as.factor(mnth),
    weekday=as.factor(weekday),
  season = fct_relevel(season, "spring", "summer", "fall", "winter"),
   holiday = fct_relevel(holiday, "no", "yes"),
   workingday = fct_relevel(workingday, "no", "yes"),
   yr = fct_relevel(yr, "2011", "2012"),
  weathersit = fct_relevel(weathersit, "clear", "mist", "light precipitation", "heavy prec
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `weathersit = fct_relevel(...)`.
Caused by warning:
! 1 unknown level in `f`: heavy precipitation
```

```r
glimpse(bike)
```

```
Rows: 731
Columns: 16
$ instant    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ dteday     <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-05~
$ season     <fct> winter, winter, winter, winter, winter, winter, winter, win~
$ yr         <fct> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,~
$ mnth       <fct> jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan,~
$ holiday    <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, no,~
$ weekday    <fct> sat, sun, mon, tue, wed, thu, fri, sat, sun, mon, tue, wed,~
$ workingday <fct> no, no, yes, yes, yes, yes, yes, no, no, yes, yes, yes, yes~
$ weathersit <fct> mist, mist, clear, clear, clear, clear, mist, mist, clear, ~
$ temp       <dbl> 0.3441670, 0.3634780, 0.1963640, 0.2000000, 0.2269570, 0.20~
$ atemp      <dbl> 0.3636250, 0.3537390, 0.1894050, 0.2121220, 0.2292700, 0.23~
$ hum        <dbl> 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0.518261,~
$ windspeed  <dbl> 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.1869000, 0.08~
$ casual     <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25, 38, 54~
$ registered <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1280, 122~
$ cnt        <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 126~
```

We get a warning that no day (during these 2 years) is classified as heavy precipitation. This seems strange to me so I went in to the hourly data and only 3 hours during the two years have heavy precipitation. I have myself not been to Washington DC and am no expert on this but my intuition makes me doubt the quality of the data here. If I would make important decisions from the data I would investigate this further and check multiple sources to see if the data is correct, or not use the variable.

```
hour<-read.csv("~/Downloads/bike+sharing+dataset/hour.csv")
sum(hour$weathersit==4)
```

```
[1] 3
```

Convert weather data to normal units

We want units to be in degrees C, humidity in %, and wind-speed in m/s. Since there were no units given in the information accompanying the dataset I had to go to http://www.freemeteo.com and look at historical data to deduce which units are used.

```
bike<-bike %>%
  mutate(
    temp = temp * 41, #temperature in degrees C
    atemp = atemp * 50, # perceived temperature in degrees C
    hum = hum * 100, #humidity in %
    windspeed = windspeed * 67 *1000/3600 #windspeed in m/s
  )
glimpse(bike)
```

```
Rows: 731
Columns: 16
$ instant    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ dteday     <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-05~
$ season     <fct> winter, winter, winter, winter, winter, winter, winter, win~
$ yr         <fct> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,~
$ mnth       <fct> jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan,~
$ holiday    <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, no,~
$ weekday    <fct> sat, sun, mon, tue, wed, thu, fri, sat, sun, mon, tue, wed,~
$ workingday <fct> no, no, yes, yes, yes, yes, yes, no, no, yes, yes, yes, yes~
$ weathersit <fct> mist, mist, clear, clear, clear, clear, mist, mist, clear, ~
$ temp       <dbl> 14.110847, 14.902598, 8.050924, 8.200000, 9.305237, 8.37826~
$ atemp      <dbl> 18.181250, 17.686950, 9.470250, 10.606100, 11.463500, 11.66~
$ hum        <dbl> 80.5833, 69.6087, 43.7273, 59.0435, 43.6957, 51.8261, 49.86~
$ windspeed  <dbl> 2.986078, 4.625587, 4.621306, 2.983287, 3.478417, 1.666908,~
$ casual     <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25, 38, 54~
$ registered <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1280, 122~
$ cnt        <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 126~
```

Rename variables

```
bike<- bike%>%
  rename(total =cnt, nonmember=casual, member=registered,tempC=temp,feeltempC=atemp)

glimpse(bike)
```

```
Rows: 731
Columns: 16
$ instant    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
$ dteday     <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-05~
$ season     <fct> winter, winter, winter, winter, winter, winter, winter, win~
$ yr         <fct> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,~
$ mnth       <fct> jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan, jan,~
$ holiday    <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, no,~
$ weekday    <fct> sat, sun, mon, tue, wed, thu, fri, sat, sun, mon, tue, wed,~
$ workingday <fct> no, no, yes, yes, yes, yes, yes, no, no, yes, yes, yes, yes~
$ weathersit <fct> mist, mist, clear, clear, clear, clear, mist, mist, clear, ~
$ tempC      <dbl> 14.110847, 14.902598, 8.050924, 8.200000, 9.305237, 8.37826~
$ feeltempC  <dbl> 18.181250, 17.686950, 9.470250, 10.606100, 11.463500, 11.66~
$ hum        <dbl> 80.5833, 69.6087, 43.7273, 59.0435, 43.6957, 51.8261, 49.86~
$ windspeed  <dbl> 2.986078, 4.625587, 4.621306, 2.983287, 3.478417, 1.666908,~
$ nonmember  <int> 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25, 38, 54~
$ member     <int> 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1280, 122~
$ total      <int> 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1321, 126~
```

Save data-set

```r
save(bike,file="~/Documents/Undervisning/IBP nytt/Case linear regression/bike.RData")
```