RLevelUp07 통계적 가설 검정

기술 통계와 추론 통계

- 기술 통계(Descriptive statistics) : 데이터를 요약해 설명하는 통계 기법
- 추론 통계(Inferential statistics) : 요약 이후의 어떤 값이 발생할 확률을 계산하는 통계 기법

통계적 가설 검정

- 유의확률을 이용해 가설을 검정하는 방법을 '통계적 가설 검정'이라고 한다.
- 유의확률은 실제로 집단간 차이가 없는데 우연한 차이가 있는 데이터가 추출될 확률을 의미

두 집단의 평균의 차이가 있는지 검정(t검정)

두 변수간의 관계가 있는지 검정하는 상관분석

01 compact 자동차와 suv자동차의 도시 연비 t검정

- 데이터 셋 : ggplot2 패키지의 mpg 데이터셋
- 소형차와 SUV가 도시 연비에서 통계적으로 유의한 차이가 있는가?

In [2]:

```
library(ggplot2)
dat <- as.data.frame(ggplot2::mpg)
dim(dat); class(dat)
```

234 11

'data.frame'

In [3]:

```
library(dplyr)
mpg_diff <- dat %>% select(class, cty) %>%
             filter(class %in% c("compact", "suv"))
head(mpg_diff)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

class	cty
compact	18
compact	21
compact	20
compact	21
compact	16
compact	18

In [4]:

```
table(mpg_diff$class)
```

```
compact     suv
     47      62
```

t.test() 이용하여 t검정을 수행하기

t-test의 유형 3가지

- 독립 표본 t-test : 서로 다른 두개의 그룹 간의 평균 비교
- 대응 표본 t-test : 하나의 집단에 대한 비교
- 단일 표본 t-test : 특정 집단의 평균이 어떤 숫자와 같은지 다른지를 비교

In [7]:

```
mpg_diff$class
```

'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact'
'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'suv' 'suv'
'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv'
'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv'
'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv'
'compact' 'compact' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv'
'compact' 'compact' 'compact' 'compact' 'suv' 'suv' 'suv' 'suv' 'suv' 'suv'
'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact'
'compact' 'compact' 'compact' 'compact' 'suv' 'suv' 'compact' 'compact' 'compact'
'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact' 'compact'
'compact' 'compact' 'compact'

서로 다른 두개의 그룹

- t.test(관측치~집단구분기준, 데이터프레임, var.equal=(T:독립표본), conf.level=0.95)

In [5]:

```
t.test(cty~class, data=mpg_diff, var.equal=T)
```

```
        Two Sample t-test

data:  cty by class
t = 11.917, df = 107, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.525180 7.730139
sample estimates:
mean in group compact     mean in group suv
             20.12766              13.50000
```

- p-value가 유의확률을 의미 5%를 판단 기준으로 삼고, p-value가 0.05미만이면 '집단 간 차이가 통계적으로 유의하다'
- p-value가 0.05보다 작기 때문에 'compact'와 'suv'간 평균 도시 연비 차이가 통계적으로 유의하다.

실습해보기 7-1

- 일반 휘발유(Regular)를 사용하는 자동차와 고급 휘발유(Premium)를 사용하는 자동차간 도시 연비 차이가 통계적으로 유의한지 알아보자.

02 상관분석 - 두 변수간의 관계성 분석

- '상관분석(Correlation Analysis)'은 두 연속 변수가 서로 관련이 있는지 검정하는 통계 분석 방법
- 상관계수는 0~1 사이의 값을 지니고 1에 가까울 수록 관련성이 크다는 것을 의미

In [11]:

```
library(dplyr)
bike <- read.csv("D:WWdatasetWWBikeWWtrain_bike.csv")
head(bike,10)
```

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	r
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	
2011-01-01 06:00:00	1	0	0	1	9.02	13.635	80	0.0000	2	
2011-01-01 07:00:00	1	0	0	1	8.20	12.880	86	0.0000	1	
2011-01-01 08:00:00	1	0	0	1	9.84	14.395	75	0.0000	1	
2011-01-01 09:00:00	1	0	0	1	13.12	17.425	76	0.0000	8	

데이터 탐색

- dim()
- head(), tail()
- summary(), str()

In [12]:

```
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 ...
```

In [13]:

```
bike <- as.data.frame(bike)
is(bike)
```

'data.frame' 'list' 'oldClass' 'vector' 'listOrNULL'

날씨(weather)와 count(렌탈대수)는 어느정도 관계가 있을까?

In [15]:

```
cor.test(bike$weather, bike$count)
```

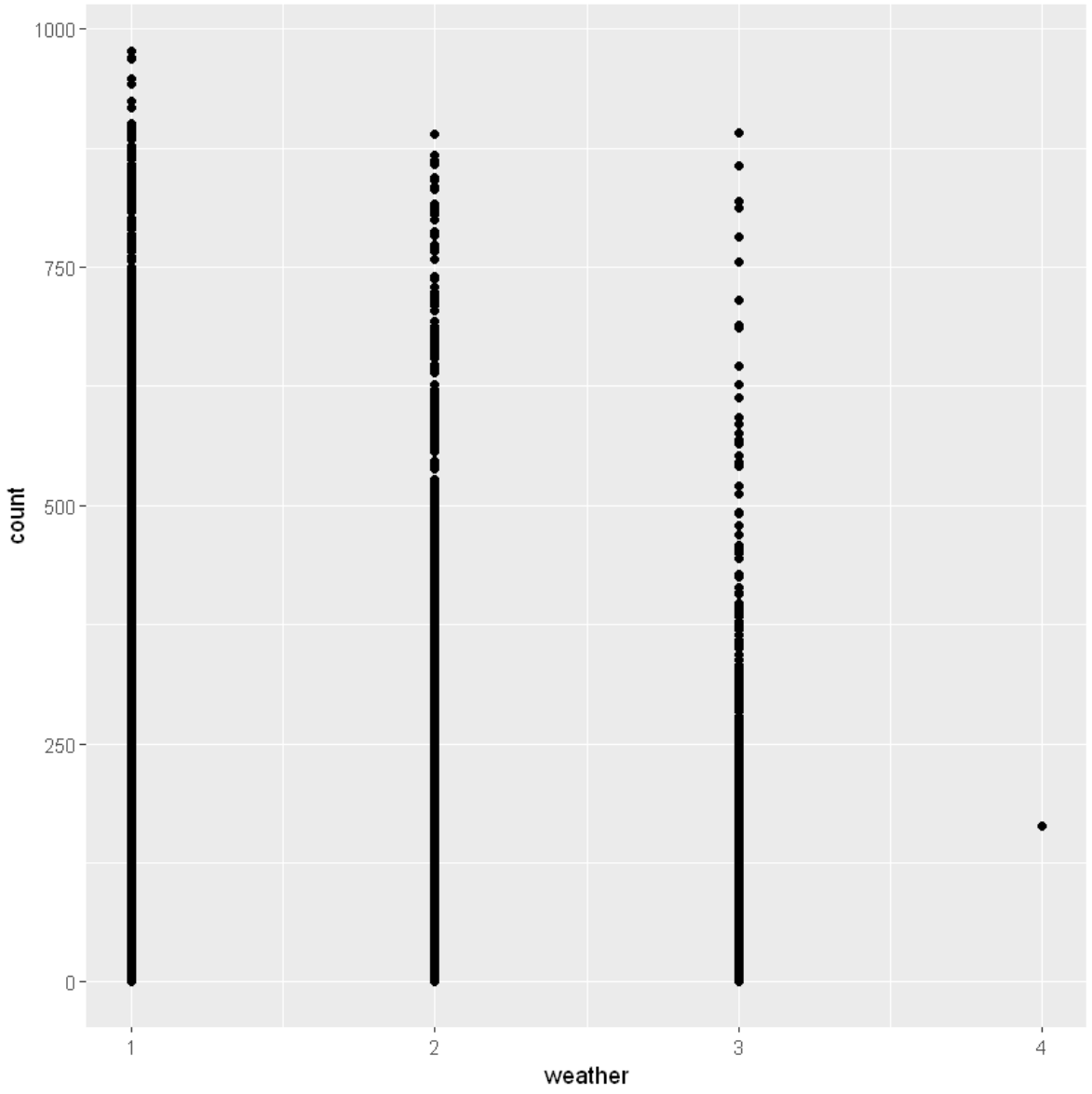
```
        Pearson's product-moment correlation

data:  bike$weather and bike$count
t = -13.535, df = 10884, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1470852 -0.1101359
sample estimates:
       cor
-0.1286552
```

- p-value가 0.05미만이므로 통계적으로 유의하다고 말할 수 있다.(상관이 있다.)
- cor이 -0.12이므로 반비례 관계

In [16]:

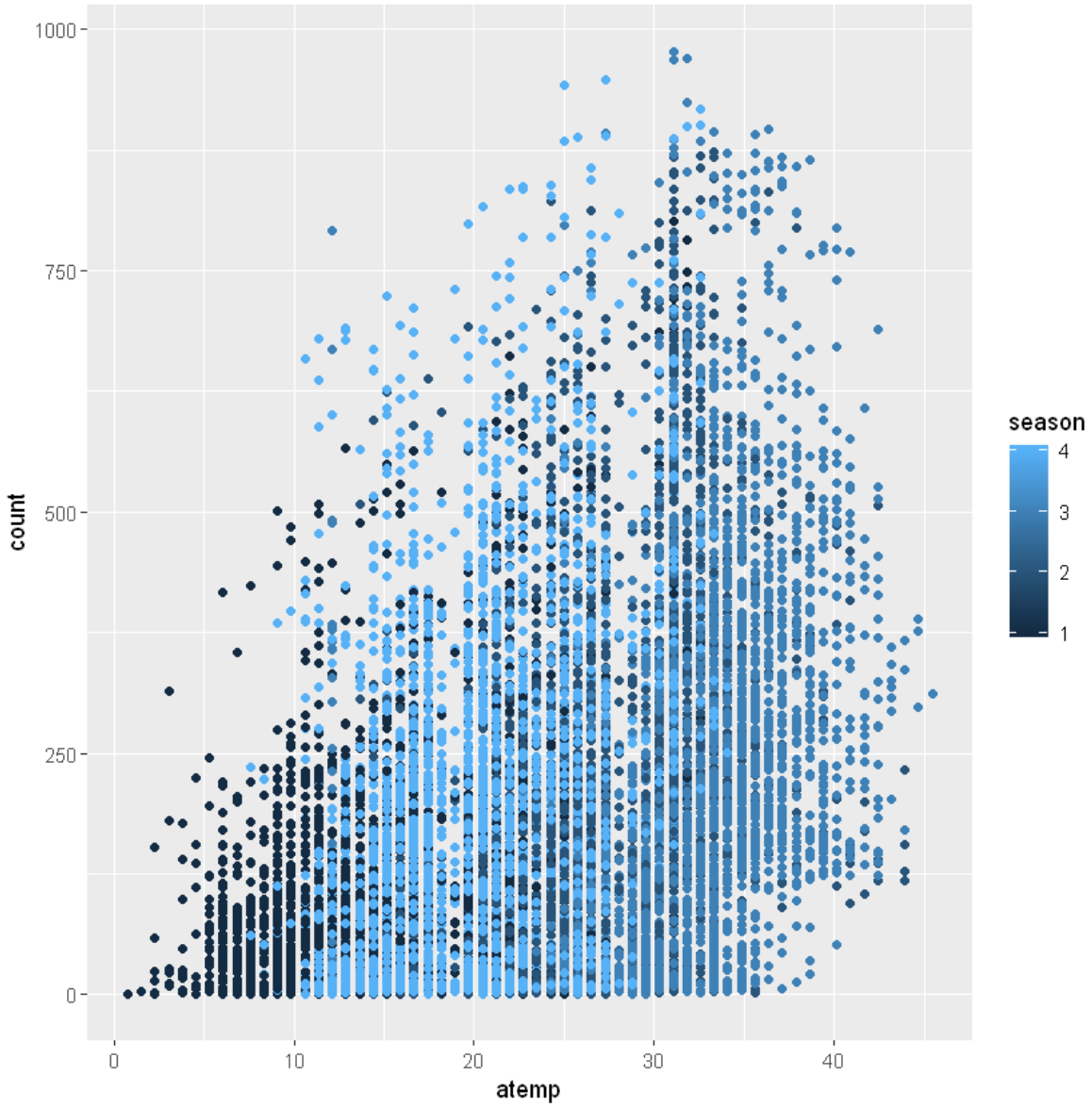
```
ggplot(data=bike, aes(x=weather, y=count)) + geom_point()
```



그렇다면 체감온도와 count는?

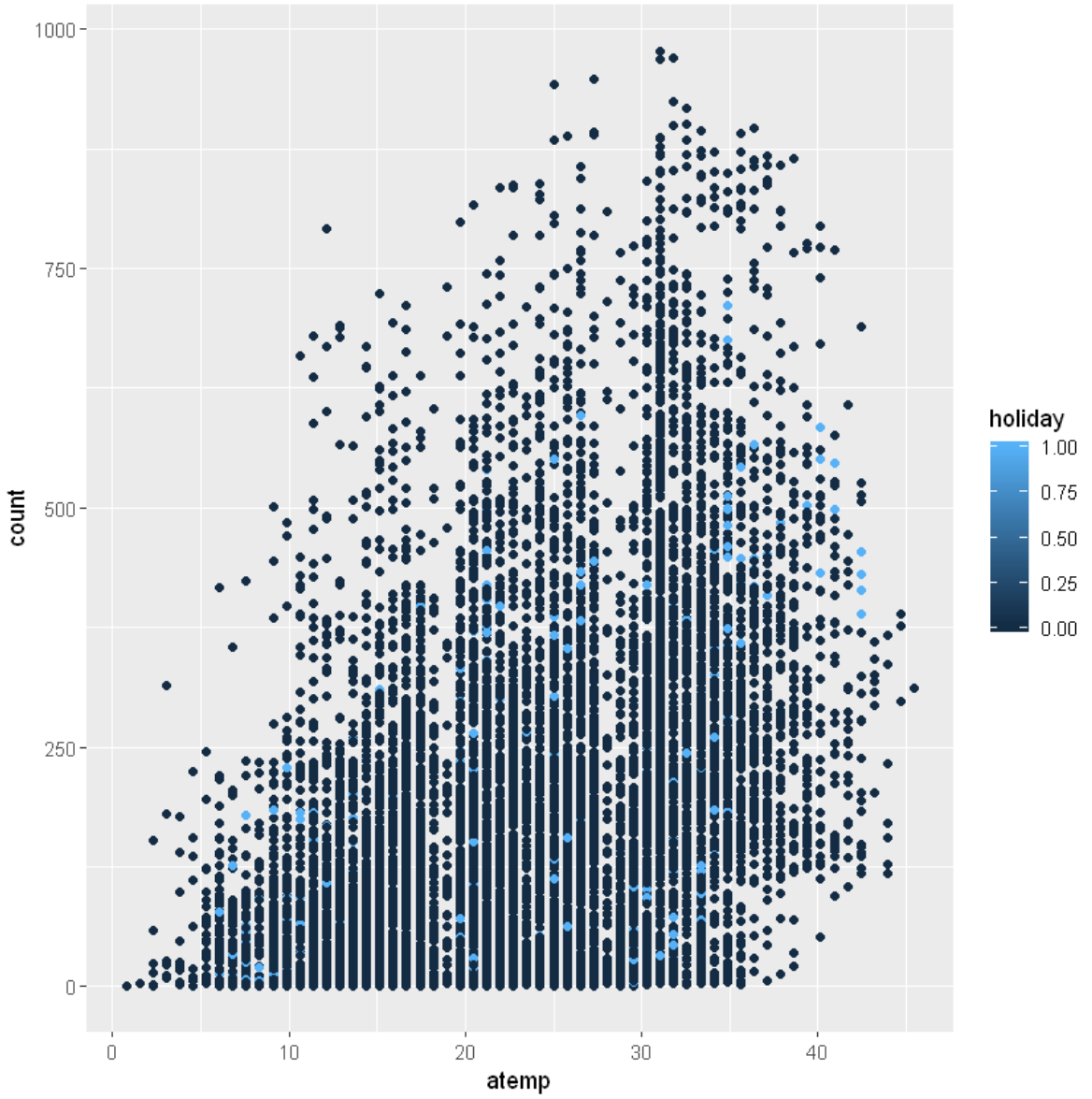
In [19]:

```
ggplot(bike, aes(x=atemp, y=count, color=season)) + geom_point()
```



In [23]:

```
ggplot(bike, aes(x=atemp, y=count, color=holiday)) + geom_point()
```



상관행렬 보기

- cor()

In [31]:

```
is(mtcars)
str(mtcars)
```

'data.frame' 'list' 'oldClass' 'vector' 'listOrNULL'

```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

In [32]:

```
is(bike)
```

'data.frame' 'list' 'oldClass' 'vector' 'listOrNULL'

In [33]:

```
cor_val <- cor(bike)
round(cor_val,2)
```

```
Error in cor(bike): 'x' must be numeric
Traceback:

1. cor(bike)
2. stop("'x' must be numeric")
```

In [34]:

```
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    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ holiday   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ workingday: num  0 0 0 0 0 0 0 0 0 0 ...
 $ weather   : num  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  : num  81 80 80 75 75 75 80 86 75 76 ...
 $ windspeed : num  0 0 0 0 0 ...
 $ casual    : num  3 8 5 3 0 0 2 1 1 8 ...
 $ registered: num  13 32 27 10 1 1 0 2 7 6 ...
 $ count     : num  16 40 32 13 1 1 2 3 8 14 ...
```

In [36]:

```
dat <- bike %>% select(-datetime)
```

In [38]:

```
cor_val <- cor(dat)
round(cor_val,2)
```

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual
season	1.00	0.03	-0.01	0.01	0.26	0.26	0.19	-0.15	0.10
holiday	0.03	1.00	-0.25	-0.01	0.00	-0.01	0.00	0.01	0.04
workingday	-0.01	-0.25	1.00	0.03	0.03	0.02	-0.01	0.01	-0.32
weather	0.01	-0.01	0.03	1.00	-0.06	-0.06	0.41	0.01	-0.14
temp	0.26	0.00	0.03	-0.06	1.00	0.98	-0.06	-0.02	0.47
atemp	0.26	-0.01	0.02	-0.06	0.98	1.00	-0.04	-0.06	0.46
humidity	0.19	0.00	-0.01	0.41	-0.06	-0.04	1.00	-0.32	-0.35
windspeed	-0.15	0.01	0.01	0.01	-0.02	-0.06	-0.32	1.00	0.09
casual	0.10	0.04	-0.32	-0.14	0.47	0.46	-0.35	0.09	1.00
registered	0.16	-0.02	0.12	-0.11	0.32	0.31	-0.27	0.09	0.50
count	0.16	-0.01	0.01	-0.13	0.39	0.39	-0.32	0.10	0.69

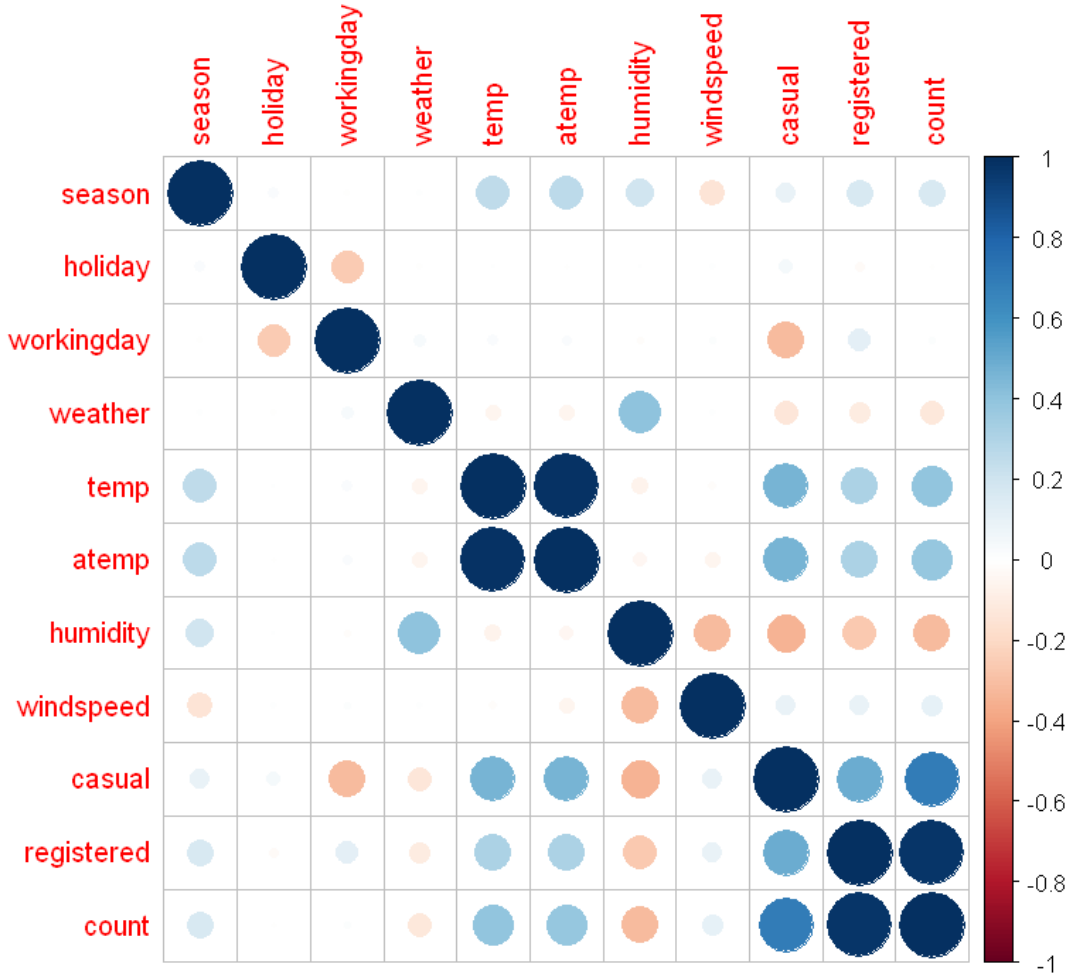
In [40]:

```
# install.packages("corrplot")
library(corrplot)
```

```
Warning message:
"unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/window
s/contrib/3.5: (http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5:)
  URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/3.5/PACKAGES'를 열 수
 없습니다"Warning message:
"package 'corrplot' is in use and will not be installed"
```

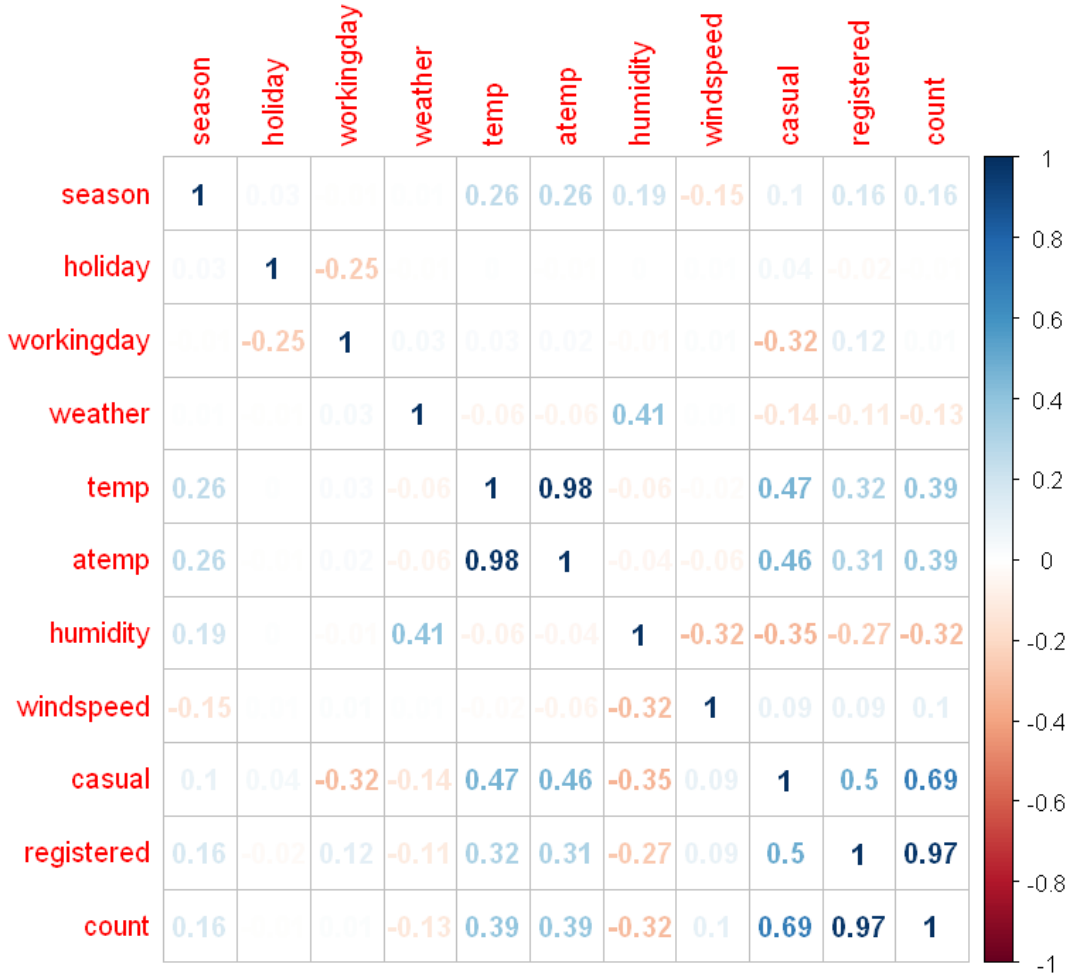
In [41]:

```
corrplot(cor_val)
```



In [42]:

```
corrplot(cor_val, method="number")
```

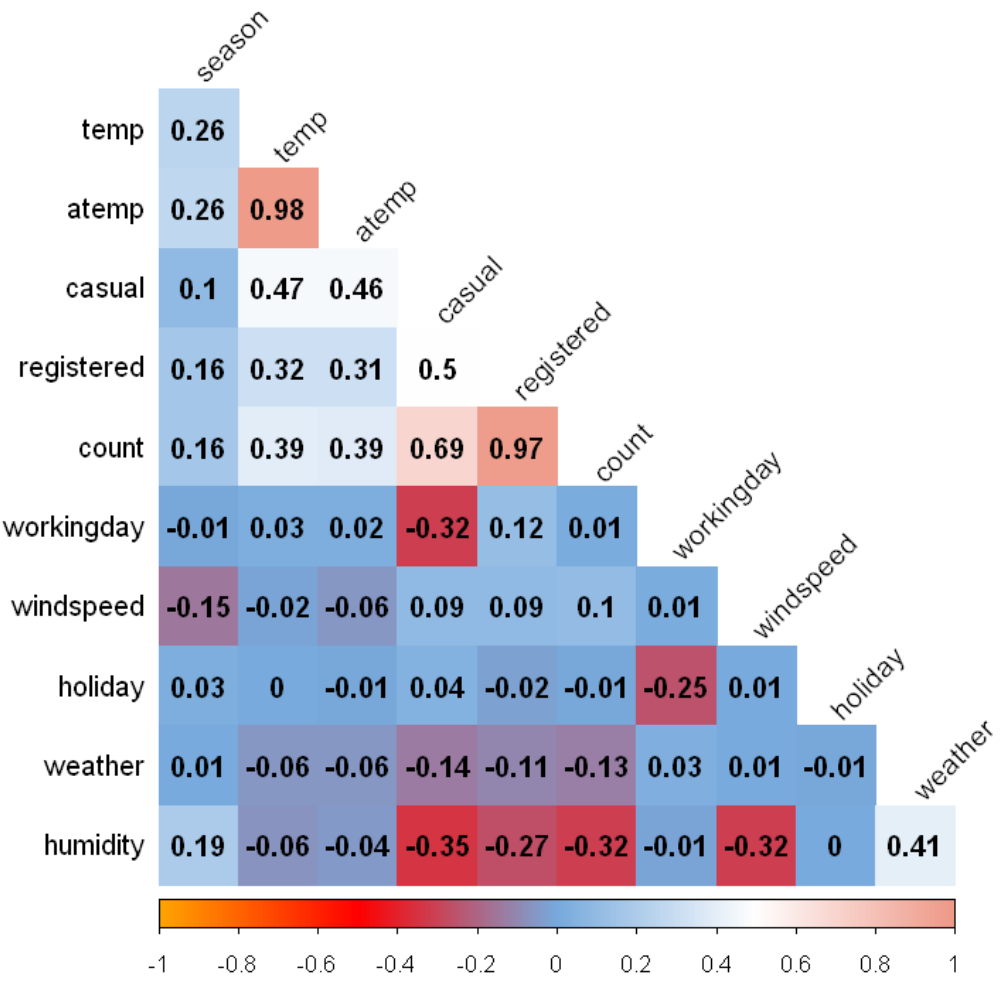


In [58]:

```
col_sel <- colorRampPalette(c("orange","red","#77AADD", "#FFFFFF", "#EE9988" ) )
```

In [59]:

```
corrplot(cor_val,
         method="color",  # 색깔로 표현
         col=col_sel(200),
         type="lower",    # 왼쪽 아래 행렬만 표시
         order="hclust",  # 유사한 상관계수끼리 군집화
         addCoef.col = "black",  # 상관계수 색깔
         tl.col = "black",       # 변수명 색깔
         tl.srt = 45,            # 변수명 45도 기울임
          diag = F )             # 대각행렬 제외
```



In [ ]:

```

```