

# PRAC2. Com realitzar la neteja i l'ànalisi de dades

Robert Carles i Marqueño i Arnau Janot Baró

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

```
# Requiriments
if (!require('dplyr')) install.packages('dplyr'); library(dplyr)
if (!require('dbscan')) install.packages('dbscan'); library(dbscan)
if (!require('gridExtra')) install.packages('gridExtra'); library(gridExtra)
if (!require('ggplot2')) install.packages('ggplot2'); library(ggplot2)
if (!require('grid')) install.packages('grid'); library(cluster)
if (!require('cluster')) install.packages('cluster'); library(cluster)
if (!require('fpc')) install.packages('fpc'); library(fpc)
if (!require('ggfortify')) install.packages('ggfortify'); library(ggfortify)
if (!require('tidyverse')) install.packages('tidyverse'); library(tidyverse)
if (!require('Stat2Data')) install.packages('Stat2Data'); library('Stat2Data')
if (!require('factoextra')) install.packages('factoextra'); library('factoextra')
if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
if (!require('polycor')) install.packages('polycor'); library('polycor')

# Funció multiplot
multiplot <- function(..., plotlist = NULL, file, cols = 1, layout = NULL) {
  require(grid)

  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  if (is.null(layout)) {
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                      ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots == 1) {
    print(plots[[1]])
  } else {
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    for (i in 1:numPlots) {
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                       layout.pos.col = matchidx$col))
    }
  }
}
```

```
}  
}  
}
```

## 1. Descripció del dataset

Perquè és important i quina pregunta/problema pretén respondre?

Els sistemes de lloguer de bicicletes (Bicing a Barcelona, Girocleta a Girona) han tingut un èxit notori als darrers anys i s'ha mostrat com una mesura indispensable cap a la reducció de l'emissió de gasos d'efecte hivernacle a les grans ciutats.

Tot i la bona rebuda per part dels consumidors, val a dir que un servei d'aquestes característiques no és senzill de gestionar. És per això que ens hem proposat identificar alguns aspectes importants per anticipar el comportament de la demanda.

Ens fem les següents preguntes:

A. Hi ha més demanda de bicicletes els caps de setmana?

B. L'estiu és l'època de l'any amb més demanda?

C. L'hora i les condicions climàtiques (temperatura, humitat, velocitat del vent) influeixen en el nombre de bicicletes llogades?

## 2. Integració i selecció

Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

La font del joc de dades és Kaggle, i es poden trobar els arxius al següent enllaç:

<https://www.kaggle.com/datasets/aguado/bike-rental-data-set-uci?resource=download>

Carreguem l'arxiu d'entrenament (train) amb el nom `bikes`.

```
bikes <- read.csv("train.csv", sep=";")
```

Revisem l'estructura original de `bikes`.

```
str(bikes)
```

```
## 'data.frame':    7689 obs. of  12 variables:  
## $ id          : int  3 4 5 7 8 9 10 11 12 13 ...  
## $ year        : int  2012 2011 2012 2011 2011 2011 2012 2011 2011 2011 ...  
## $ hour        : int  23 8 2 20 17 19 23 22 14 13 ...  
## $ season      : int  3 3 1 3 3 2 2 3 3 1 ...  
## $ holiday     : int  0 0 0 0 0 0 0 0 0 0 ...  
## $ workingday  : int  0 0 1 1 1 1 1 1 1 0 ...  
## $ weather     : int  2 1 1 3 3 2 2 1 1 2 ...  
## $ temp        : num  23.8 27.9 20.5 25.4 26.2 ...  
## $ atemp       : num  27.3 31.8 24.2 28.8 28.8 ...  
## $ humidity    : int  73 57 59 83 89 39 78 94 53 72 ...  
## $ windspeed   : num  11 0 0 20 0 ...  
## $ count       : int  133 132 19 58 285 326 75 160 134 94 ...
```

Hi ha **12** variables i **7689** registres.

- **ID** Número identificador. (Primary Key)
- **YEAR** Any (2011 o 2012)
- **HOURL** Hora del dia (de 0 a 23)
- **SEASON** Estació climàtica (1 = hivern, 2 = primavera, 3 = estiu, 4 = tardor)
- **HOLIDAY** Si el dia és festiu
- **WORKINGDAY** Si el dia és laboral (ni festiu ni cap de setmana)
- **WEATHER** quatre categories de millor (1) a pitjor (4)
- **TEMP** Temperatura en graus Celsius
- **ATEMP** Sensació tèrmica en graus Celsius
- **HUMIDITY** Humitat relativa
- **WINDSPEED** velocitat del vent (km/h)
- **COUNT** total de bicicletes llogades en aquella franja temporal

Si ens hi fixem, totes les variables apareixen interpretades com a numèriques. No obstant, n'hi ha dues binàries (**holiday** i **workingday**) i dues categòriques (**season** i **weather**) que han estat codificades.

Fem les modificacions pertinents per a que el dataset compleixi els requisits pertinents.

```
# Guardem el joc de dades original
bikes_original <- bikes

# Variables binàries
bikes$holiday <- as.factor(bikes$holiday)
bikes$workingday <- as.factor(bikes$workingday)

# Variables categòriques [opcional: millor per a les visualitzacions]
seasons <- c('Hivern', 'Primavera', 'Estiu', 'Tardor')
weathers <- c('Molt Bo', 'Bo', 'Dolent', 'Molt Dolent')

for (i in 1:4) {
  bikes$season[bikes$season == i] <- seasons[i]
  bikes$weather[bikes$weather == i] <- weathers[i]
}

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

Ara sí, el nostre joc de dades conté **8** variables numèriques, **2** binàries i **2** categòriques.

### 3. Neteja de les dades

Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

Comprovem que les dades no contiguin valors NA o buits.

```
# Valors NA
colSums(is.na(bikes))
```

```
##      id      year      hour      season      holiday workingday      weather
##      0         0         0         0         0           0           0
##      temp      atemp      humidity      windspeed      count
##      0         0         0         0         0
```

```
# Valors buits
colSums(bikes=="")
```

```
##      id      year      hour      season      holiday workingday      weather
##      0         0         0         0         0           0           0
##      temp      atemp      humidity      windspeed      count
##      0         0         0         0         0
```

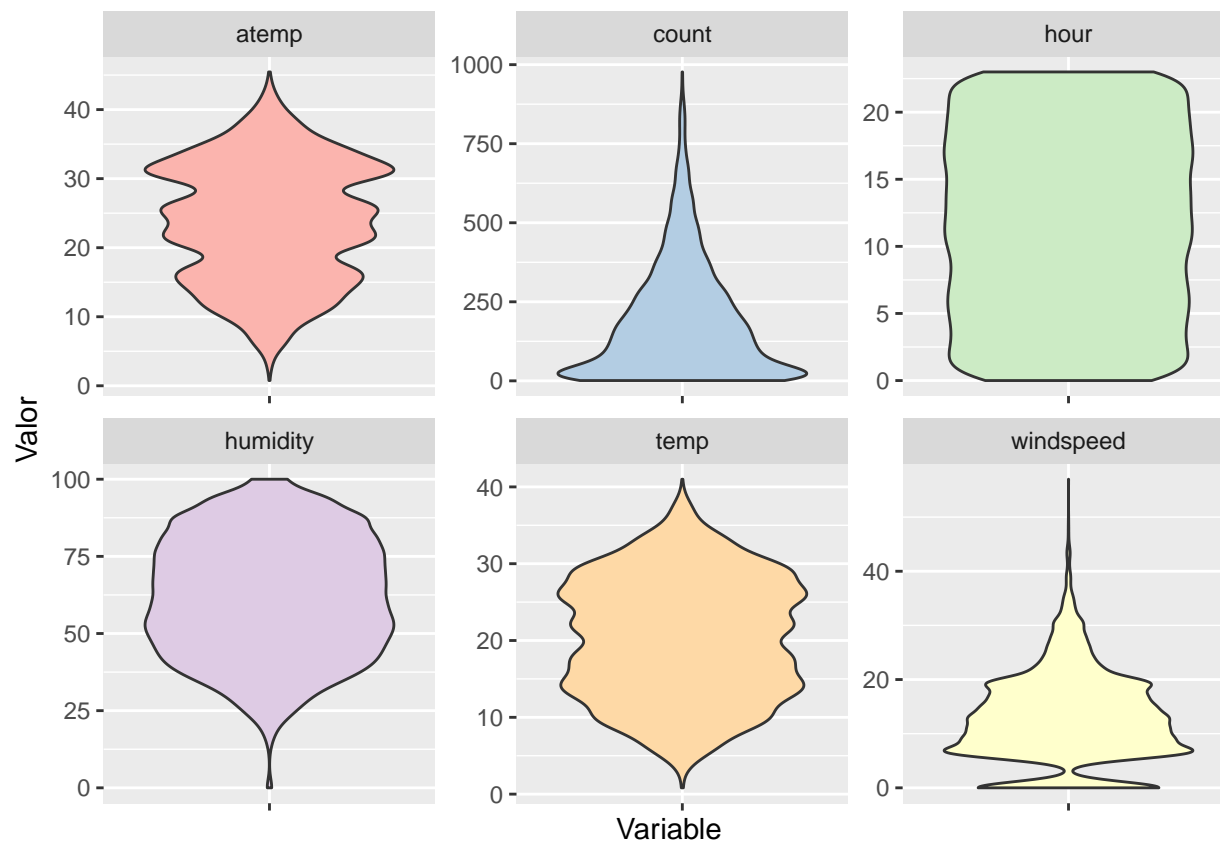
Les dades estan netes.

Fem una primera aproximació a les dades.

```
summary(bikes)
```

```
##      id      year      hour      season      holiday
## Min.   : 3      Min.   :2011      Min.   : 0.00      Estiu   :1943      0:7466
## 1st Qu.:2771     1st Qu.:2011     1st Qu.: 6.00      Hivern  :1901     1: 223
## Median :5477     Median :2011     Median :12.00     Primavera:1920
## Mean   :5463     Mean   :2011     Mean   :11.57     Tardor   :1925
## 3rd Qu.:8186     3rd Qu.:2012     3rd Qu.:18.00
## Max.   :10886     Max.   :2012     Max.   :23.00
## workingday      weather      temp      atemp      humidity
## 0:2481      Bo      :1981      Min.   : 0.82      Min.   : 0.76      Min.   : 0.00
## 1:5208      Dolent : 586      1st Qu.:13.94     1st Qu.:16.66     1st Qu.: 46.00
##           Molt Bo:5122      Median :20.50     Median :24.24     Median : 62.00
##           Mean   :20.27     Mean   :23.70     Mean   : 61.77
##           3rd Qu.:26.24     3rd Qu.:31.06     3rd Qu.: 77.00
##           Max.   :41.00     Max.   :45.45     Max.   :100.00
##      windspeed      count
## Min.   : 0.000      Min.   : 1.0
## 1st Qu.: 7.002      1st Qu.:41.0
## Median :12.998      Median :145.0
## Mean   :12.802      Mean   :191.4
## 3rd Qu.:16.998      3rd Qu.:283.0
## Max.   :56.997      Max.   :977.0
```

Mostrem la distribució de les variables numèriques.



Mostrem la distribució de les variables categòriques i numèriques.

```
# Variables categòriques
categoric_var<-c("season", "weather", "holiday", "workingday")
plotList <- list()

for(i in 1:length(categoric_var)){
  # Taula recompte
  hawk_cat <- table(bikes[categoric_var[i]])

  # Dataframe per visualitzar
  data <- data.frame(
    category=names(hawk_cat),
    count=round(as.numeric(hawk_cat)*100/sum(hawk_cat),digits = 2)
  )
  data$fraction <- data$count / sum(data$count)
  data$ymax <- cumsum(data$fraction)
  data$ymin <- c(0, head(data$ymax, n=-1))
  data$labelPosition <- (data$ymax + data$ymin) / 2
  data$label <- paste0(data$category, "\n value: ", data$count)

  # Gràfica
  ggp_geom1 <- ggplot(data, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) +
    geom_rect() +
    geom_label( x=3.5, aes(y=labelPosition, label=label), size=3) +
    ggtitle(categoric_var[i]) +
```

```

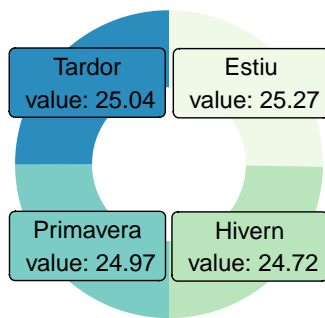
scale_fill_brewer(palette=4) +
coord_polar(theta="y") +
xlim(c(2, 4)) +
theme_void() +
theme(legend.position = "none")

# L'afegim a la plotlist
plotList[[i]] <- ggp_geom1
}

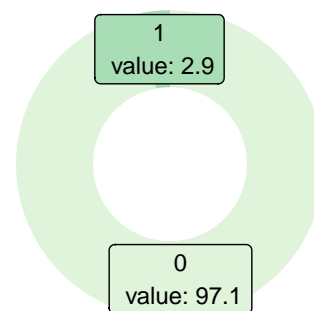
multiplot(plotlist = plotList, cols = 2)

```

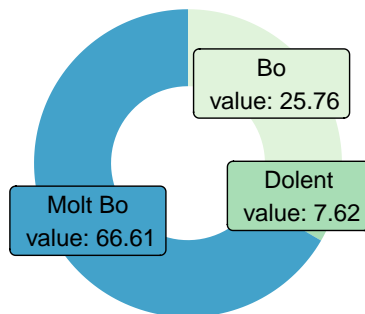
season



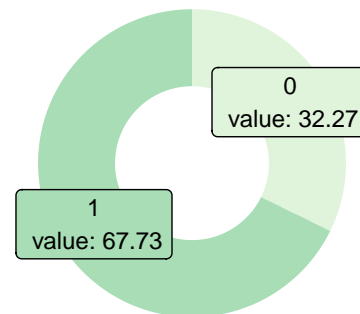
holiday



weather



workingday



## Observacions inicials

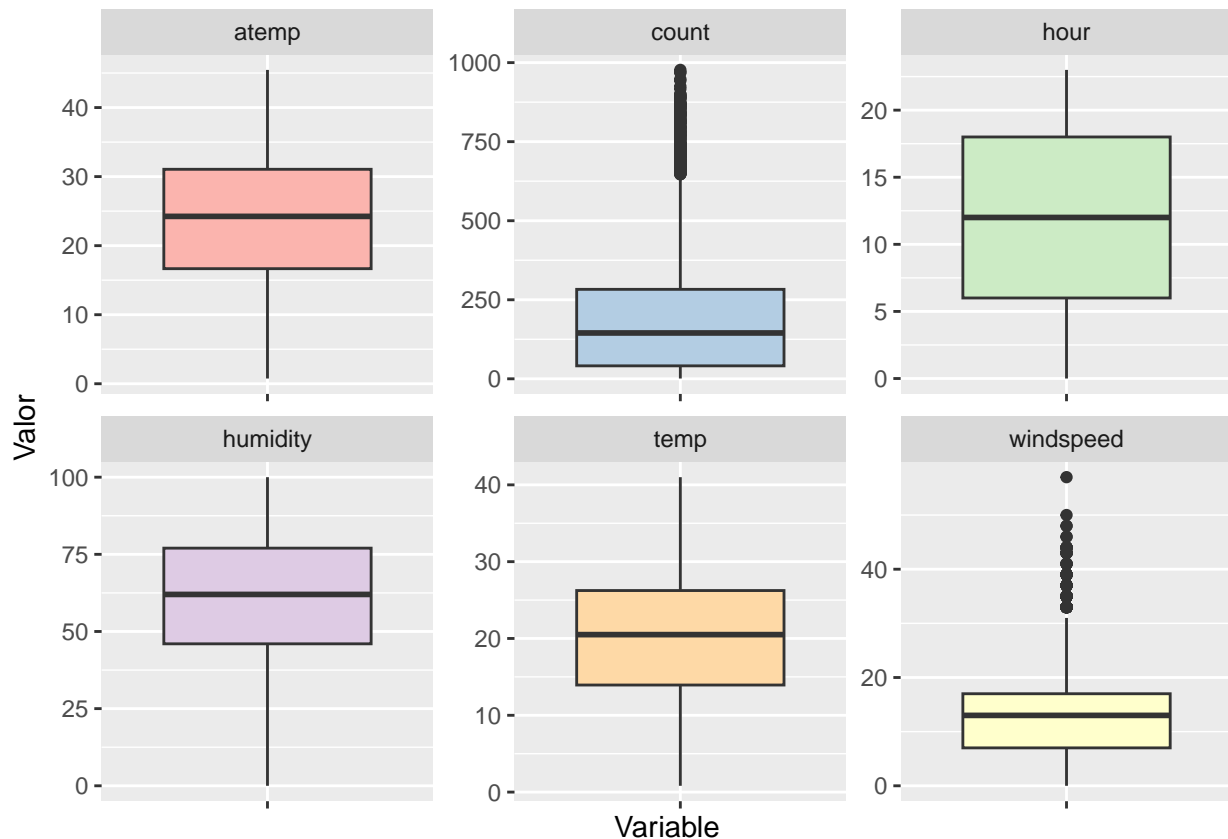
- **id** La clau primària. Els indicadors de mesura central no són d'interès. El nombre màxim (10886) és superior el nombre de registres (7689), senyal que el dataset original s'ha seccionat en dos de forma aleatòria.
- **year** Les dades pertanyen al 2011 i 2012.
- **hour** + **season** Ambdues variables estan repartides equitativament.
- **holiday** + **workingday** Com ja sabem, s'observen més dies no festius que festius (2.9%); també més laborals (67.73%) que no laborals.
- **weather** No apareix cap registre amb un temps molt dolent. La majoria d'ells ha tingut molt bones condicions climàtiques (66.61%).

- **temp + atemp** La temperatura mitjana és de 20.27°C, mentre que la sensació tèrmica és de 23.7°C.
- **humidity** La humitat mitjana és de 61.77 i la seva distribució es concentra a la franja de 35-80.
- **windspeed** La velocitat del vent ronda els 13 km/h i la seva distribució apunta a que deuen haver-hi alguns valors atípics.
- **count** De mitjana es lloguen 191 bicicletes per cada lot de temps. Mentre que el menor registre ha estat una sola bicicleta, el rècord n'han estat 977. Donada la diferència entre mitjana i mediana, valdrà la pena revisar la presència d'outliers. S'arriba a la mateixa conclusió observant el gràfic de violí.

Identifica i gestiona els valors extrems.

Revisem la presència de **valors atípics**.

```
# Boxplot ggplot2
ggplot(gather(bikes[numeric_var1],key="Variable", value="Valor"),aes(x=Variable, y=Valor, fill=Variable)) +
  geom_boxplot() +
  theme(legend.position="none", axis.text.x = element_blank()) +
  scale_fill_brewer(palette="Pastel1") +
  facet_wrap(~Variable, scales="free")
```



```
# OUTLIERS
# COUNT
outliers_count <- boxplot.stats(bikes$count)$out
length(outliers_count)
```

```
## [1] 219
```

```
# WINDSPEED
outliers_windspeed <- boxplot.stats(bikes$windspeed)$out
length(outliers_windspeed)
```

```
## [1] 154
```

Hi ha **219** outliers a `count` i `windspeed` en té **154**.

Enlloc de prescindir de tot el registre, imputarem el valor que es troba a la punta del bigoti superior del diagrama de caixa. És a dir, imputarem el *nou valor màxim* un cop borrats els valors superiors a aquest.

```
# IMPUTACIÓ DE VALORS

# COUNT
bikes[bikes$count %in% outliers_count, "count"] <- NA
bikes[is.na(bikes$count), "count"] <- max(na.omit(bikes$count))

# WINDSPEED
bikes[bikes$windspeed %in% outliers_windspeed, "windspeed"] <- NA
bikes[is.na(bikes$windspeed), "windspeed"] <- max(na.omit(bikes$windspeed))
```

## 4. Anàlisi de les dades

### 4.1. Selecció dels grups de dades

Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

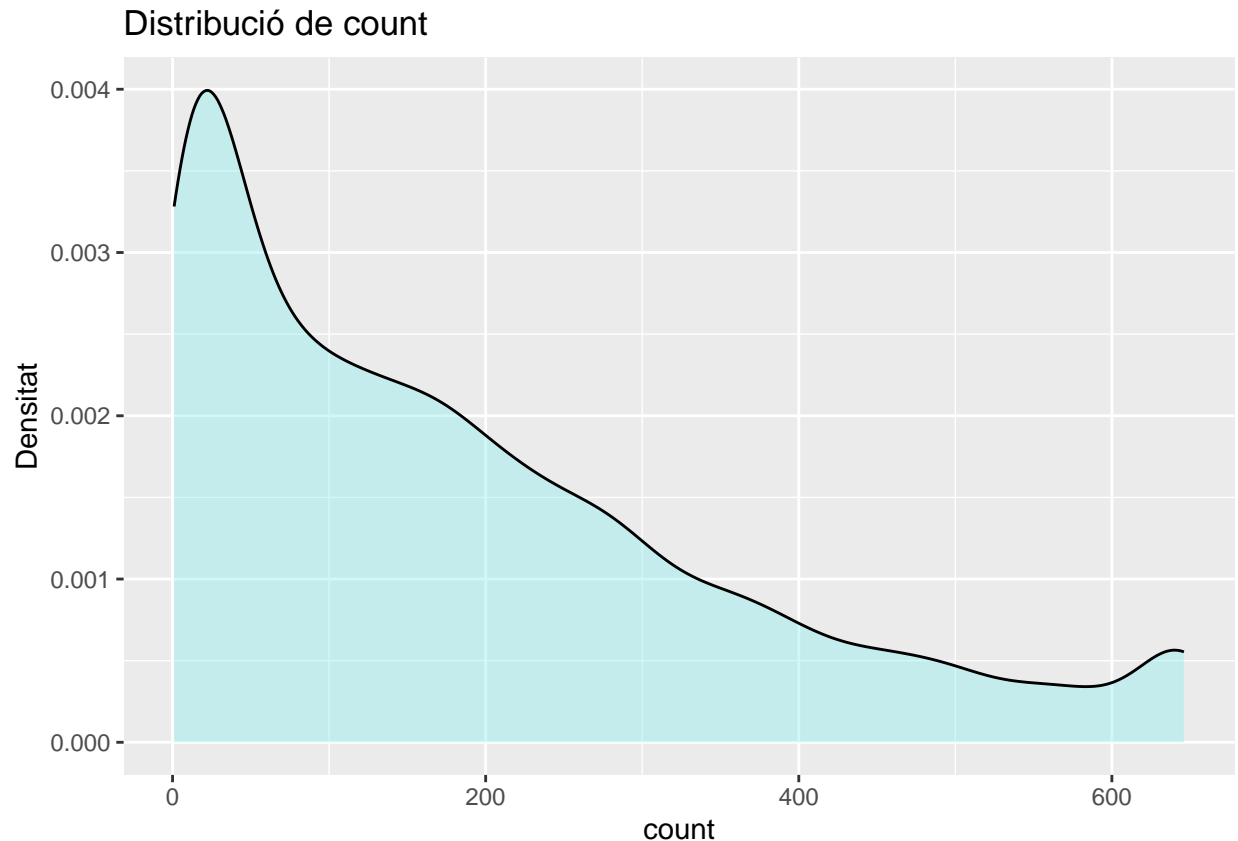
Per a respondre a les preguntes A i B s'emprarà la variable `count` i, respectivament, `workingday` i `season`.

L'última qüestió (C), en canvi, requereix un estudi holístic de les variables. Per tant, començarem fent un estudi de les **correlacions** entre les diferents variables per a identificar problemes de colinearitat alhora que observem les variables que més influeixen en el nombre de bicicletes llogades. Amb aquest subconjunt de dades, provarem de generar un **model de regressió lineal múltiple** i quantificarem el seu ajust.

### 4.2. Comprovació de la normalitat i homogeneïtat de la variància

```
ggplot(bikes, aes(x=count)) +
  geom_density(fill="darkslategray2", alpha=0.4) +
  ggtitle("Distribució de count") +
  xlab('count') +
  ylab('Densitat')
```





La distribució de count **no és normal**. Tot i haver tractat outliers, s'observa un pic desplaçat cap a l'esquerra i unes dades molt disperses. En tot cas, si tenim en compte el **teorema del límit central** (TLC) i sabem que tenim més de 30 mostres (bastantes més), podem assumir que la **mitjana de count es distribueix de forma normal**.

- L'homogeneïtat de les dades. Pregunta A.

```
# Subdataframes
bikes_work <- bikes[bikes$workingday == 1,]
bikes_fest <- bikes[bikes$workingday == 0,]

# Test d'igualtat de variàncies, 95
var.test(bikes_work$count, bikes_fest$count, conf.level = 0.95)
```

```
##
## F test to compare two variances
##
## data: bikes_work$count and bikes_fest$count
## F = 1.0389, num df = 5207, denom df = 2480, p-value = 0.2712
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9705829 1.1111785
## sample estimates:
## ratio of variances
##          1.038931
```

Donat que  $p > 0.05$ , **acceptem la igualtat de variàncies** entre els dies laborals i caps de setmana.

- L'homogeneïtat de les dades. Pregunta B.

```
# Subdataframes
bikes_Summer <- bikes[bikes$season == "Estiu",]
bikes_NotSummer <- bikes[bikes$season != "Estiu",]

# Test d'igualtat de variàncies, 95
var.test(bikes_Summer$count, bikes_NotSummer$count, conf.level = 0.95)

##
## F test to compare two variances
##
## data: bikes_Summer$count and bikes_NotSummer$count
## F = 1.2294, num df = 1942, denom df = 5745, p-value = 1.49e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.143913 1.323131
## sample estimates:
## ratio of variances
##          1.229446
```

Donat que  $p < 0.05$ , **rebutgem la igualtat de variàncies** entre l'estiu i la resta d'estacions.

## 4.3. Aplicació de proves estadístiques

### 4.3.1 Contrasts d'hipòtesi

- A) Hi ha més demanda de bicicletes els caps de setmana?

#### 1. Pregunta de recerca

La demanda de bicicletes és significativament superior el cap de setmana?

#### 2. Hipòtesi nul · la i l'alternativa

$$H_0: \mu_{capde} = \mu_{laboral}$$

$$H_1: \mu_{capde} > \mu_{laboral}$$

#### 3. Test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals

```
t.test(bikes_work$count, bikes_fest$count, var.equal = TRUE, alternative = "greater", conf.level = 0.95)

##
## Two Sample t-test
##
## data: bikes_work$count and bikes_fest$count
## t = 0.93526, df = 7687, p-value = 0.1748
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
## -2.994514      Inf
## sample estimates:
## mean of x mean of y
## 189.6709 185.7251
```

Hem obtingut un valor  $p > 0.05$ . No podem rebutjar la hipòtesi nul·la.

#### 4. Conclusió

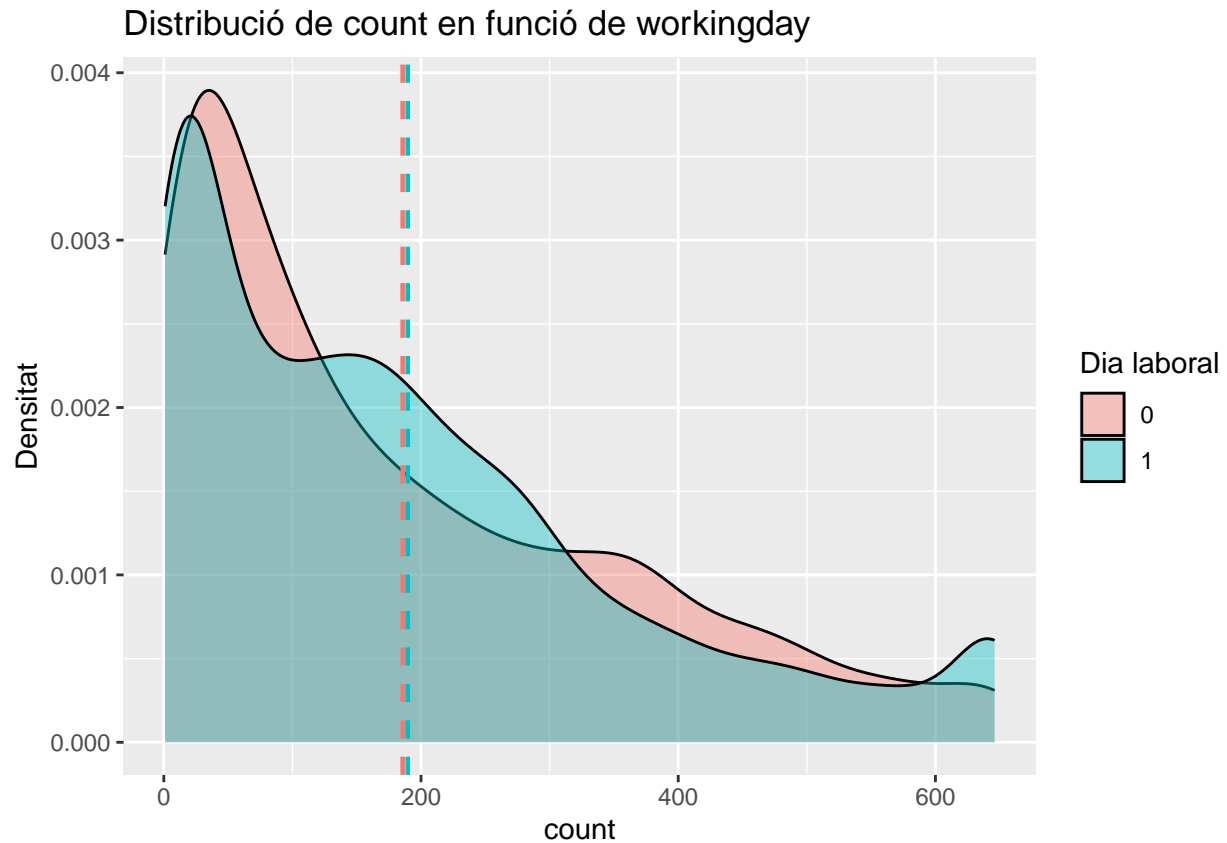
La demanda no és significativament superior durant els caps de setmana.

```
# Mitjanes de count per workingday
mean_count_df <- bikes %>%
  group_by(workingday) %>%
  summarize(mean=mean(count))

# Gràfica count + workingday
ggplot(bikes, aes(x=count, fill=workingday)) +
  geom_density(alpha=0.4) +
  scale_fill_discrete(name = "Dia laboral") +
  ggtitle("Distribució de count en funció de workingday") +
  xlab('count') +
  ylab('Densitat')+
  geom_vline(data = mean_count_df, aes(
    xintercept = mean, color = workingday),linetype = "dashed", size=0.8)+
  guides(color = FALSE, size = FALSE)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
## Warning: The '<scale>' argument of 'guides()' cannot be 'FALSE'. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



- B) Hi ha més demanda de bicicletes a l'estiu?

### 1. Pregunta de recerca

La demanda de bicicletes és significativament superior el cap de setmana?

### 2. Hipòtesi nul·la i l'alternativa

$$H_0: \mu_{estiu} = \mu_{altre}$$

$$H_1: \mu_{estiu} > \mu_{altre}$$

### 3. Test de dues mostres independents sobre la mitjana amb variàncies desconegudes diferents

```
t.test(bikes_Summer$count, bikes_NotSummer$count, var.equal = FALSE, alternative = "greater", conf.level = 0.95)
```

```
##
## Welch Two Sample t-test
##
## data: bikes_Summer$count and bikes_NotSummer$count
## t = 11.531, df = 3078.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  46.76934      Inf
## sample estimates:
## mean of x mean of y
##  229.1657  174.6121
```

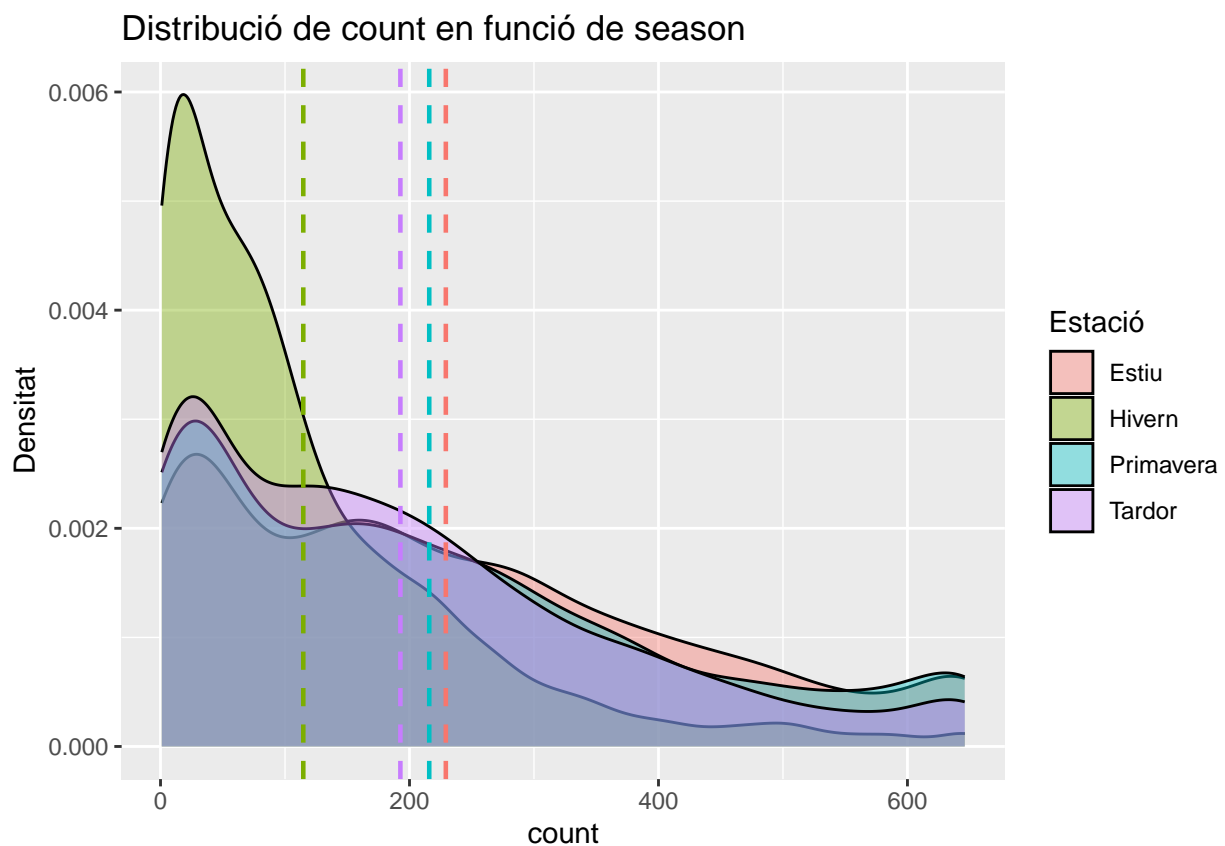
Hem obtingut un valor  $p < 0.05$ . Rebutgem la hipòtesi nul·la.

#### 4. Conclusió

L'estiu és l'estació de l'any que registra una demanda significativament superior a la resta.

```
# Mitjanes de count per season
mean_count_df <- bikes %>%
  group_by(season) %>%
  summarize(mean=mean(count))

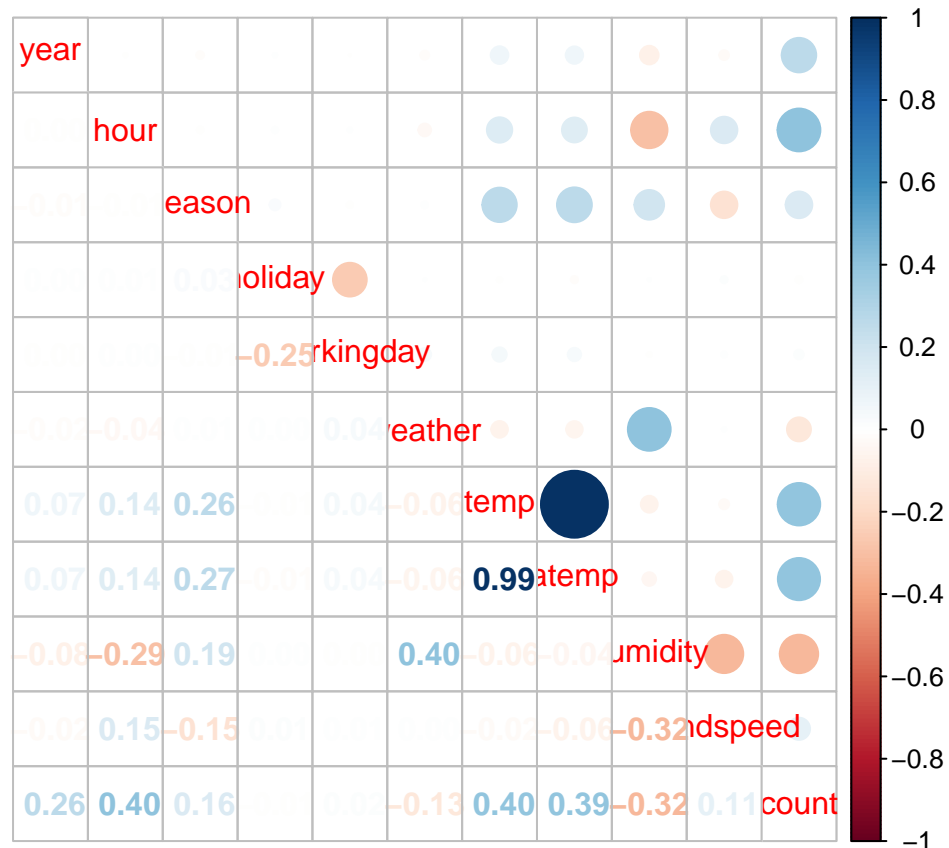
# Gràfica count + season
ggplot(bikes, aes(x=count, fill=season)) +
  geom_density(alpha=0.4) +
  scale_fill_discrete(name = "Estació") +
  ggtitle("Distribució de count en funció de season") +
  xlab('count') +
  ylab('Densitat') +
  geom_vline(data = mean_count_df, aes(
    xintercept = mean, color = season), linetype = "dashed", size=0.8) +
  guides(color = FALSE, size = FALSE)
```



#### 4.3.2 Anàlisi de correlacions

Estudiem la correlació i colinealitat de les variables:

```
corr_df <- bikes_original[2:12]
corr <- hetcor(corr_df, use="complete.obs")
corrplot.mixed(corr=corr$correlations)
```

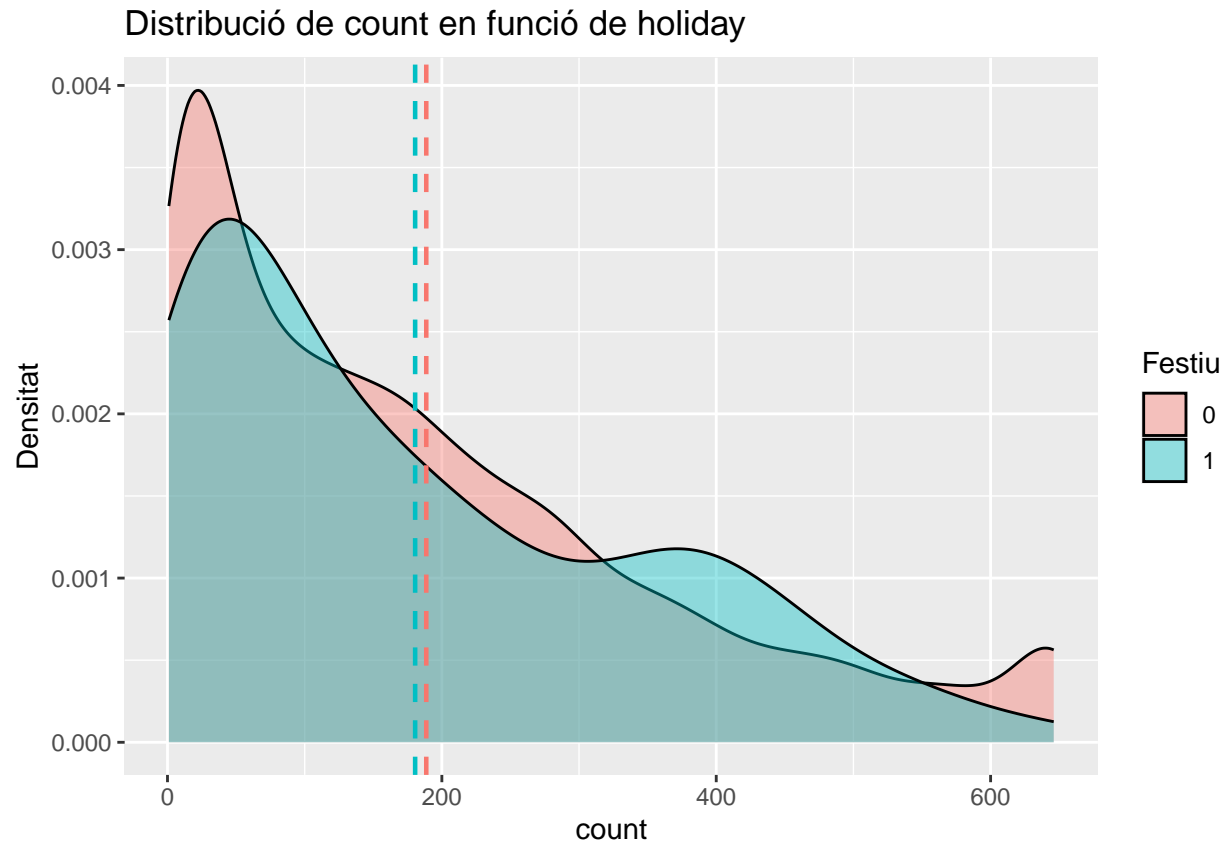


## Observacions

- No s'ha detectat cap correlació entre la festivitat o *laboralitat* i el nombre de bicis llogades. Tampoc és massa significativa la variable *windspeed* ni *season*.

```
# Mitjanes de count per holiday
mean_count_df <- bikes %>%
  group_by(holiday) %>%
  summarize(mean=mean(count))

# Gràfica count + holiday
ggplot(bikes, aes(x=count, fill=holiday)) +
  geom_density(alpha=0.4) +
  scale_fill_discrete(name = "Festiu") +
  ggtitle("Distribució de count en funció de holiday") +
  xlab('count') +
  ylab('Densitat') +
  geom_vline(data = mean_count_df, aes(
    xintercept = mean, color = holiday), linetype = "dashed", size=0.8) +
  guides(color = FALSE, size = FALSE)
```



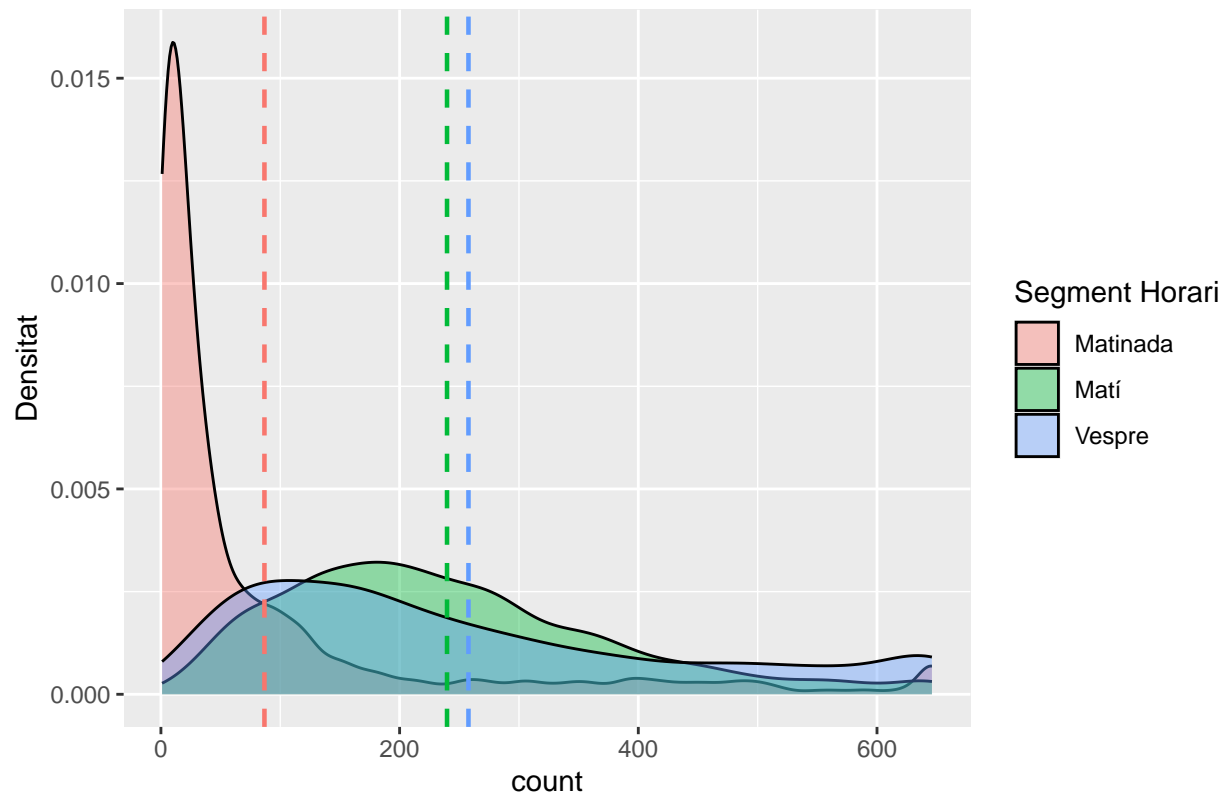
- Les variables que fan augmentar **count** de forma més notòria són l'**hora** i la **temperatura/sensació tèrmica** - seguits de l'**any**. En canvi, a mesura que augmenta la **humitat** disminueix la demanda.

```
# Discretitzem la variable hora
bikes$segment_horari <- cut(bikes$hour, breaks = c(-1, 8, 16, 23),
                           labels = c("Matinada", "Matí", "Vespre"))

# Mitjanes de count per segment_horari
mean_count_df <- bikes %>%
  group_by(segment_horari) %>%
  summarize(mean=mean(count))

# Gràfica count + segment_horari
ggplot(bikes, aes(x=count, fill=segment_horari)) +
  geom_density(alpha=0.4) +
  scale_fill_discrete(name = "Segment Horari") +
  ggtitle("Distribució de count en funció de segment_horari") +
  xlab('count') +
  ylab('Densitat') +
  geom_vline(data = mean_count_df, aes(
    xintercept = mean, color = segment_horari), linetype = "dashed", size=0.8) +
  guides(color = FALSE, size = FALSE)
```

Distribució de count en funció de segment\_horari

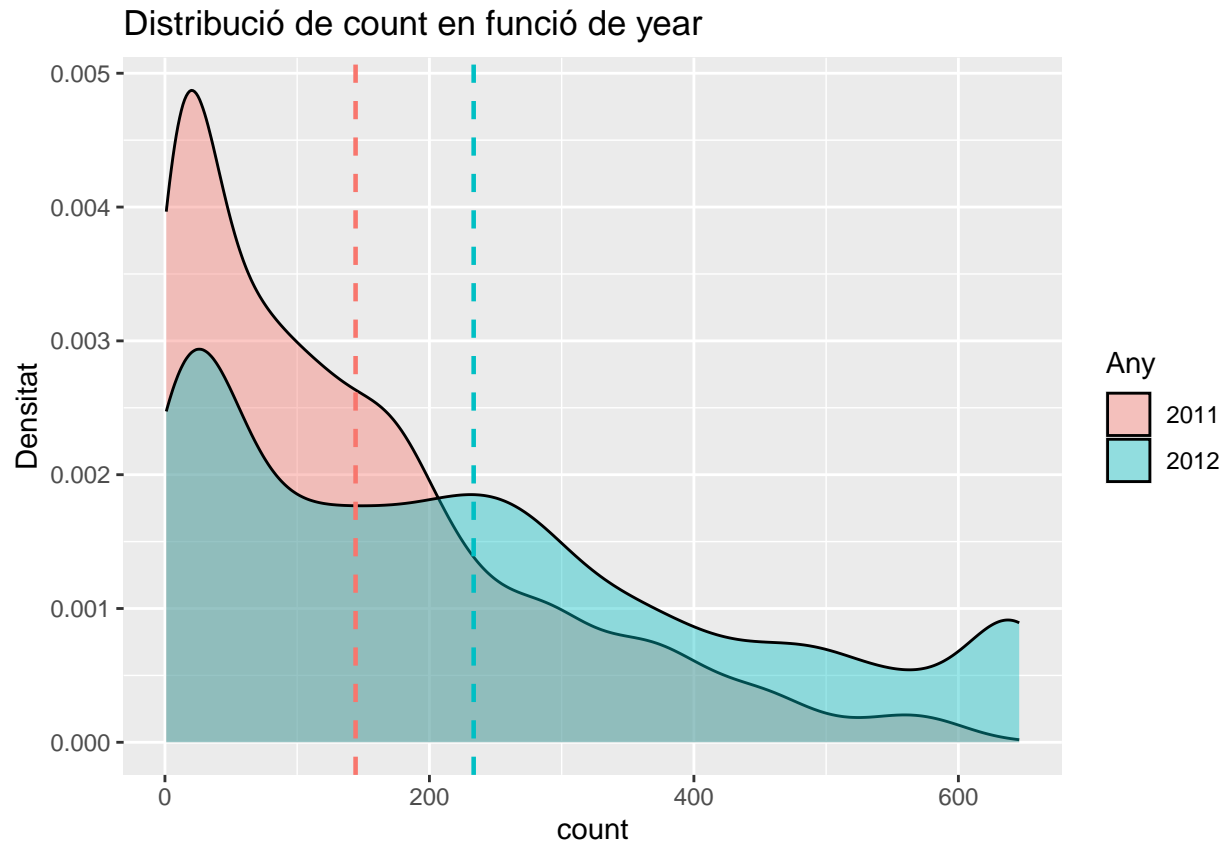


```
# Al haver només dos anys, el visualitzem igual
bikes$year <- as.factor(bikes$year)

# Mitjanes de count per year
mean_count_df <- bikes %>%
  group_by(year) %>%
  summarize(mean=mean(count))

# Gràfica count + year
ggplot(bikes, aes(x=count, fill=year)) +
  geom_density(alpha=0.4) +
  scale_fill_discrete(name = "Any") +
  ggtitle("Distribució de count en funció de year") +
  xlab('count') +
  ylab('Densitat') +
  geom_vline(data = mean_count_df, aes(
    xintercept = mean, color = year), linetype = "dashed", size=0.8) +
  guides(color = FALSE, size = FALSE)
```





La demanda és major al 2012. Unes dades molt prometedores de cara al negoci!

- La major correlació es troba entre la temperatura i la sensació tèrmica (el 99%). Per tant, podrem prescindir d'alguna d'elles.

### 4.3.3 Regressió lineal múltiple

Per últim, estimarem per mínims quadrats ordinaris un model lineal que expliqui la variable `count` en funció de les variables `year`, `hour`, `temp` i `humidity`:

```
model <- lm(count~year+hour+temp+humidity, data = bikes)
summary(model)
```

```
##
## Call:
## lm(formula = count ~ year + hour + temp + humidity, data = bikes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -308.88  -92.10  -24.83   60.08  568.75
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.23822    7.89227   3.198  0.00139 **
## year2012     76.51509    3.11037  24.600 < 2e-16 ***
```

```
## hour          7.73457    0.23590  32.787 < 2e-16 ***
## temp          7.31309    0.20018  36.532 < 2e-16 ***
## humidity     -1.81998    0.08405 -21.655 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 135.6 on 7684 degrees of freedom
## Multiple R-squared:  0.3855, Adjusted R-squared:  0.3852
## F-statistic: 1205 on 4 and 7684 DF,  p-value: < 2.2e-16
```

## Observacions

- Donat un p-valor inferior a 2.2e-16, podem afirmar que **totes les variables escollides són significatives**. Aquest resultat, de fet, és coherent amb els resultats de l'apartat anterior.
- El coeficient de determinació indica que **només un 38.55% de la variància** de les observacions **queda explicada** pel model lineal.

## 5. Conclusions finals

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Per acabar el projecte podem extreure les següents conclusions finals.

Podem afirmar que hi ha una **demanda significativament superior de bicicletes durant l'estiu**. Podem associar aquest resultat al bon temps. El bon temps fa que la gent vulgui estar més al carrer realitzant activitats a l'aire lliure com pot ser utilitzar les bicicletes.

També podem afirmar que **no hi ha una diferència significativa en la demanda de bicicletes entre els dies laborals i els caps de setmana**. Per tant ni el treball ni les vacances son doncs patrons que puguin justificar l'ús de la bicicleta.

L'**hora** del dia, la **temperatura/sensació tèrmica**, l'**humitat** i l'**any** són les variables més rellevants per predir la demanda de bicicletes ja que tenen una **influència significativa** en el nombre de bicicletes llogades.

El que ens ha sorprès molt també és la **correlació negativa de count i humitat**: a mesura que augmenta la humitat, la demanda de bicicletes disminueix.

El **model de regressió lineal múltiple** utilitzat per predir la demanda de bicicletes és significatiu i **explica aproximadament el 38.55% de la variabilitat observada**. Per tant cal remarcar que el 38,55% queda cobert per les variables estudiades però la resta queda en mans d'altres factors influents en la demanda que no hem estudiat aquí.

Finalment, els resultats obtinguts tenen un alt valor afegit per a les empreses que proporcionen serveis de bicing ja que els permet implementar una estratègia de negoci eficaç.