Tarea4_5

Jorge Porras Araya

02/11/2019

R Markdown

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.6.1
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```
## Warning: package 'purrr' was built under R version 3.6.1
```

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

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.1
```

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

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

```
glimpse(flights)
```

```
## Observations: 336,776
## Variables: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,...
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 55...
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 60...
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2...
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 7...
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 7...
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -...
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV",...
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79...
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN...
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR"...
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL"...
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138...
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 94...
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013...
```

```
##glimpse(airports)
```

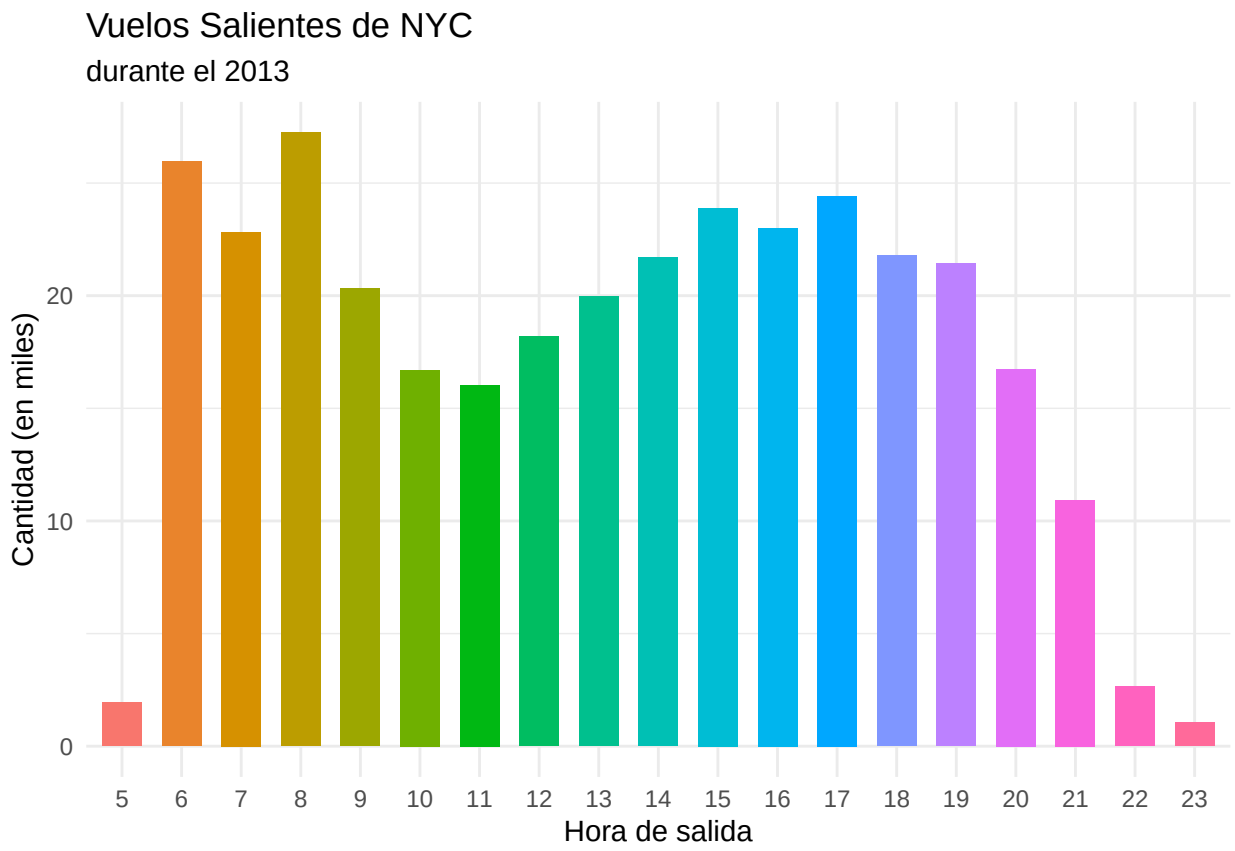
Primera Parte

A) Flights es una tabla de una base de datos que contiene información de los vuelos que han salido de cualquiera de los 3 aeropuertos de la ciudad de Nueva York en el año 2013 (hacia cualquier destino).

B) A continuación un gráfico de barras que representa la sumatoria de vuelos salientes de los aeropuertos de Nueva York a cada hora del día.

```
Cantidad_hora <- data.frame(table(flights$hour))
colnames(Cantidad_hora) = c("Hora", "Cantidad")
Cantidad_hora <- Cantidad_hora %>% filter(Cantidad > 100)

gv <- ggplot(data = Cantidad_hora,
             aes(x = Hora,
                 y = Cantidad / 1000,
                 width = 0.666,
                 fill = Hora)) +
  geom_bar(stat = "identity") +
  labs(title = "Vuelos Salientes de NYC",
       subtitle = "durante el 2013",
       y = "Cantidad (en miles)",
       x = "Hora de salida") +
  theme_minimal() +
  theme(legend.position = "none")
gv
```



Se puede observar que a las 8 de la mañana es cuando mas vuelos parten de Nueva York y de medianoche a las 4am (inclusive) practicamente no hay salidas (con la excepción de que existió una partida a la 1am pero se elimina del gráfico pues no es apreciable dada las grandes diferencias de magnitudes)

C) Gráfico de barras horizontal

Las Aerolineas más utilizadas, salientes de Nueva York se pueden observar a continuación, donde la más importante es United Air Lines.

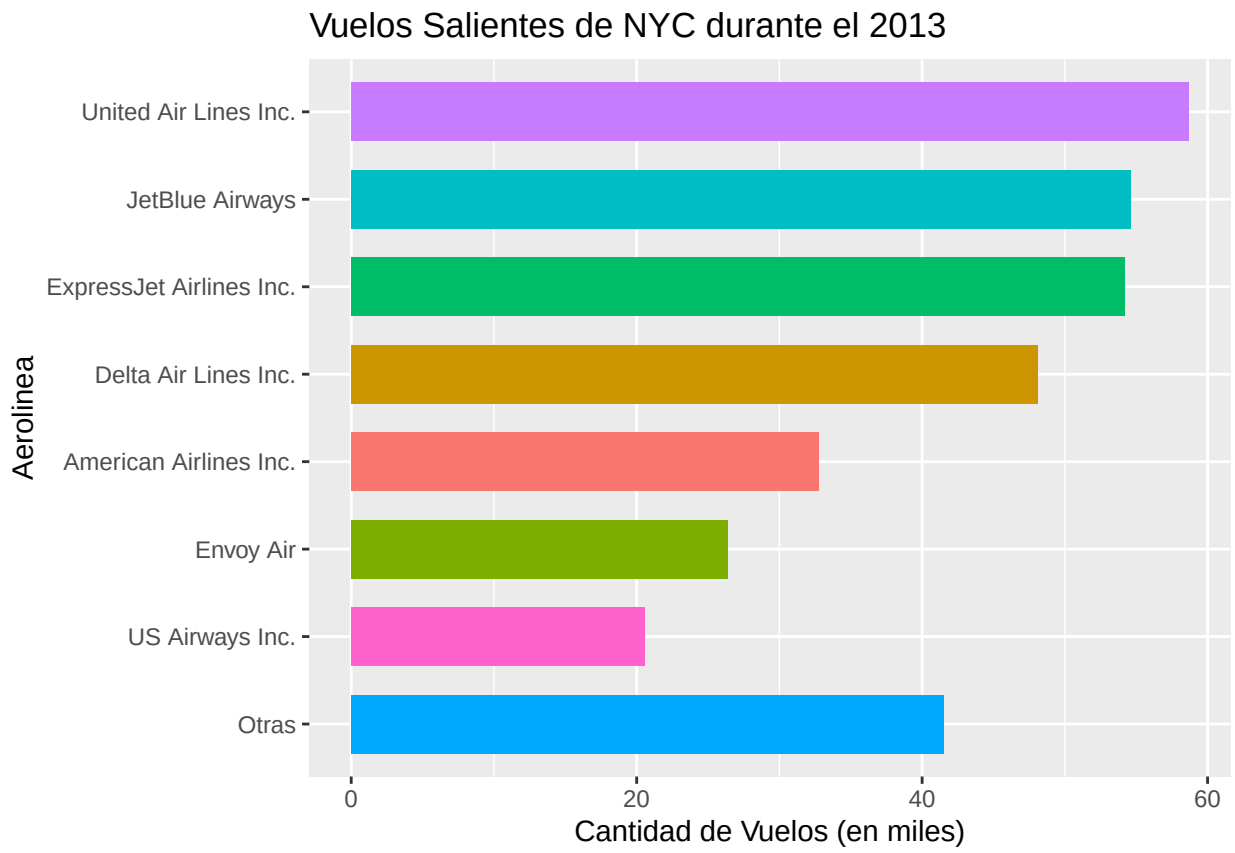
```
tf <- data.frame(table(flights$carrier))
colnames(tf) = c("carrier", "cantidad")
tf <- merge(x = tf, y = airlines, all.x = TRUE)
tf <- tf %>% arrange(desc(cantidad))
tf <- tf[,c(2, 3)]
t1_a = tf[1:7,]
t1_b = tf[8:nrow(tf),]
t1_c <- data.frame("name" = "Otras",
                   "cantidad" = sum(t1_b$cantidad))

t1 = rbind(t1_a, t1_c)
t1$order <- seq(1, length(t1$cantidad))

t1 = t1 %>% mutate(porc = round(100 * cantidad / sum(cantidad), 2))
colnames(t1) = c("Cantidad", "Aerolinea","Orden", "Porcentaje")
```

```r
gh <- ggplot(data = t1,
             aes(x = reorder(Aerolinea, -Orden),
                 y = Cantidad / 1000,
                 width = 0.666,
                 fill = Aerolinea)) +
  geom_bar(stat = "identity") +
  labs(title = "Vuelos Salientes de NYC durante el 2013",
       x = "Aerolinea",
       y = "Cantidad de Vuelos (en miles)") +
  theme(legend.position = "none") +
  coord_flip()

gh
```



Para obtener los nombres de las aerolineas se tiene que

D) Gráfico de Pastel

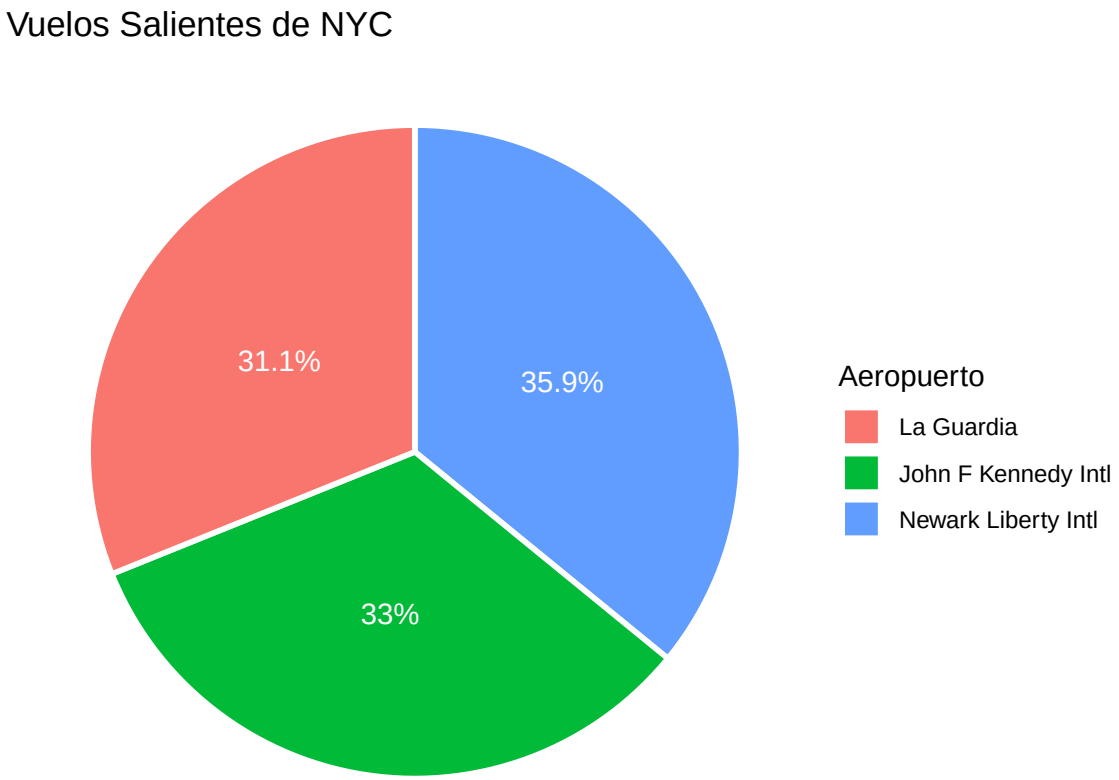
Dado que se tienen 3 aeropuertos, a continuacion se muestra el porcentaje de vurlos salientes desde cada uno de los 3 aeropuertos.

```r
porc_aeropuerto <- data.frame(table(flights$origin))
colnames(porc_aeropuerto) = c("faa", "cantidad")
porc_aeropuerto <- merge(x = porc_aeropuerto, y = airports, all.x = TRUE)
porc_aeropuerto <- porc_aeropuerto[,c(1,2,3)]
```

```r
porc_aeropuerto <- porc_aeropuerto %>% mutate(porc = cantidad / sum(porc_aeropuerto$cantidad)) %>% arra
porc_aeropuerto$porc <- round(porc_aeropuerto$porc * 100 , 1)
porc_aeropuerto <- ungroup(porc_aeropuerto)
colnames(porc_aeropuerto) = c("faa", "Cantidad", "Aeropuerto", "Porcentaje")
porc_aeropuerto$order <- seq(1, length(porc_aeropuerto$Cantidad))

gp <- ggplot(data = porc_aeropuerto,
             aes(x = "",
                 y = Porcentaje,
                 fill = reorder(Aeropuerto, -order))) +
  geom_bar(width = 1,
           size = 1,
           color = "white",
           stat = "identity") +
  geom_text(aes(label = paste0(Porcentaje, "%")),
            position = position_stack(vjust = 0.5),
            color = "white") +
  labs(title = "Vuelos Salientes de NYC", fill = "Aeropuerto")  +
  coord_polar("y", direction = 1) +
  theme_void()

gp
```



II Parte

1) Binomial

Suponga que el 30% de los estudiantes del ITCR son mujeres. Si se toma una muestra de 10 mujeres al azar.

a) Escriba la fórmula de la distribución que modela esta situación.

La ecuacion que describe la distribucion binomial es:

$$P_{(x)} = \binom{n}{x} p^x q^{n-x}$$

donde:

$$C_{n,x} = \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

b) ¿Cuál es la probabilidad de que en la muestra haya al menos 4 mujeres?

```
p <- 0.3
q <- 1-p
n <- 10
x <- 4
```

Como debe de haber 4 o mas entonces se debe realizar el acumulado de probabilidad de 4 a 10.

```
sum(dbinom(seq(4,n), n, p))
```

```
## [1] 0.3503893
```

ó realizar el acumulado de 0, 1, 2, 3 y restarlo de 1.

```
1- sum(dbinom(c(0,1,2,3), n, p))
```

```
## [1] 0.3503893
```

ó realizar el acumulado con la funcion pbinom y restarlo de 1.

```
1- pbinom(3, n, p)
```

```
## [1] 0.3503893
```

c) ¿Cuál es la probabilidad de que en la muestra haya más de 4 y a lo sumo 8 mujeres?

La probabilidad de que haya más de 4 y a lo sumo 8

```
p <- 0.3
q <- 1-p
n <- 10
x <- c(5, 6, 7, 8)
```

```
sum(dbinom(x, n, p))
```

```
## [1] 0.1501246
```

d) ¿Cuál es la probabilidad de que haya exactamente 5 mujeres en la muestra?

La probabilidad de que haya exactamente 5 mujeres

```
p <- 0.3
q <- 1-p
n <- 10
x <- 5
```

```
dbinom(x, n, p)
```

```
## [1] 0.1029193
```

e) Determine la media y la desviación estándar que se esperaría ver en la muestra.

```
p <- 0.3
q <- 1-p
n <- 10
```

La media se obtine de asi:

```
media <- n*p
media
```

```
## [1] 3
```

y la varianza:

```
varianza <- n*p*q
varianza
```

```
## [1] 2.1
```

2) Hipergeométrica

Constantemente, la gente que posee un vehículo vive quejándose de las fallas mecánicas que estos presentan, tanto, porque es un gasto tanto para sus bolsillos como por las implicaciones en la disponibilidad de uso. Popularmente, se tiene la creencia de que una persona podría olvidarse de este tipo de imprevistos si tuviese la capacidad de pago para adquirir un vehículo nuevo de agencia, situación que no es tan cierta. Suponga que en Costa Rica, la compañía de autos A debe hacer efectiva su garantía para algunos usuarios que adquirieron de agencia, uno de sus autos modelo X. Sabiendo que para el mes de enero pasado, disponía de 30 vehículos de dicho modelo X en sus agencias donde 3 de ellos venía con un desperfecto de fábrica y que durante ese mes logró vender 8 de esos autos, realice lo siguiente:

a) Verifique que se cumple la condición para que sea hipergeométrica y escriba la ecuación de la distribución que modela esta situación

Se cumple la condición pues con cada vehículo vendido ya se tendria uno menos de donde escoger (además es un valor pequeño que es lo que la hipergeometrica pide).

La ecuación es la siguiente:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

b) ¿Cuál es la probabilidad de que ninguno de los autos vendidos requiera la garantía durante el primer año?.

La probabilidad de que de los 8 autos vendidos, 8 salieran buenos es de:

```
m <- 27
n <- 3
k <- 8
x <- 8
```

```
dhyper(x, m, n, k)
```

```
## [1] 0.3793103
```

c) Determine la probabilidad de que al menos dos de los autos vendidos requiera hacer uso de la garantía.

```
m <- 27
n <- 3
k <- 8
x <- seq(0,5)
```

```
1-sum(dhyper(x, m, n, k))
```

```
## [1] 0.9862069
```

3) Geométrica

Un juego consiste en gira runa ruleta que tiene 3 opciones de premio y 13 opciones no premio. Una persona que juegue la ruleta ganará si al girarla obtiene la opción "premio" antes del cuarto intento.

a) Escriba la fórmula de la distribución que modela esta situación

La ecuacion es la siguiente:

$$f(x) = \begin{cases} p(1-p)^{x-1} & \text{cuando} \quad x \epsilon \{1, \quad 2, \quad 3, \quad \dots\} \\ 0 & \text{de lo contrario} \end{cases}$$

b) Calcule la probabilidad de si una persona gane en una jugada en la ruleta.

```
q <- 2
premio <- 3
npremio <- 13
prob <- premio/(premio + npremio)
```

Entonces la probabilidad de gane en una jugada es de:

```
pgeom(q,prob)
```

```
## [1] 0.463623
```