Tarea Examen 1

Yimmy Eman

2022-07-09

Pregunta 1

Intenta describir con frases entendibles el conjunto de vuelos retrasados. Intenta dar afirmaciones como por ejemplo:

- Un vuelo tiende a salir unos 20 minutos antes el 50% de las veces y a salir tarde el 50% de las veces restantes.
- Los vuelos de la compañía XX llegan siempre 20 minutos tarde
- El 95% de los vuelos a HNL llegan a tiempo, pero el 5% restante se retrasan más de 3 horas. Intenta dar por lo menos 5 afirmaciones verídicas en base a los datos que tenemos disponibles.

Planteamiento 1:

- El top 3 de los vuelos más lento pertenecen a las compañías US, B6 y 9E.
- El top 3 de los vuelos más rápidos pertenecen a las compañías DL y EV.

```
slow.flights <- group_by(flights,carrier) %>%
  summarise(speed = distance/ air_time * 60) %>%
  arrange(speed)

fast.flights <- group_by(flights,carrier) %>%
  summarise(speed = distance/ air_time * 60) %>%
  arrange(desc(speed))

head(slow.flights, 3)
```

```
## # A tibble: 3 x 2
## # Groups:   carrier [3]
##   carrier speed
##   <chr>   <dbl>
## 1 US       76.8
## 2 B6       84.7
## 3 9E       92.5
```

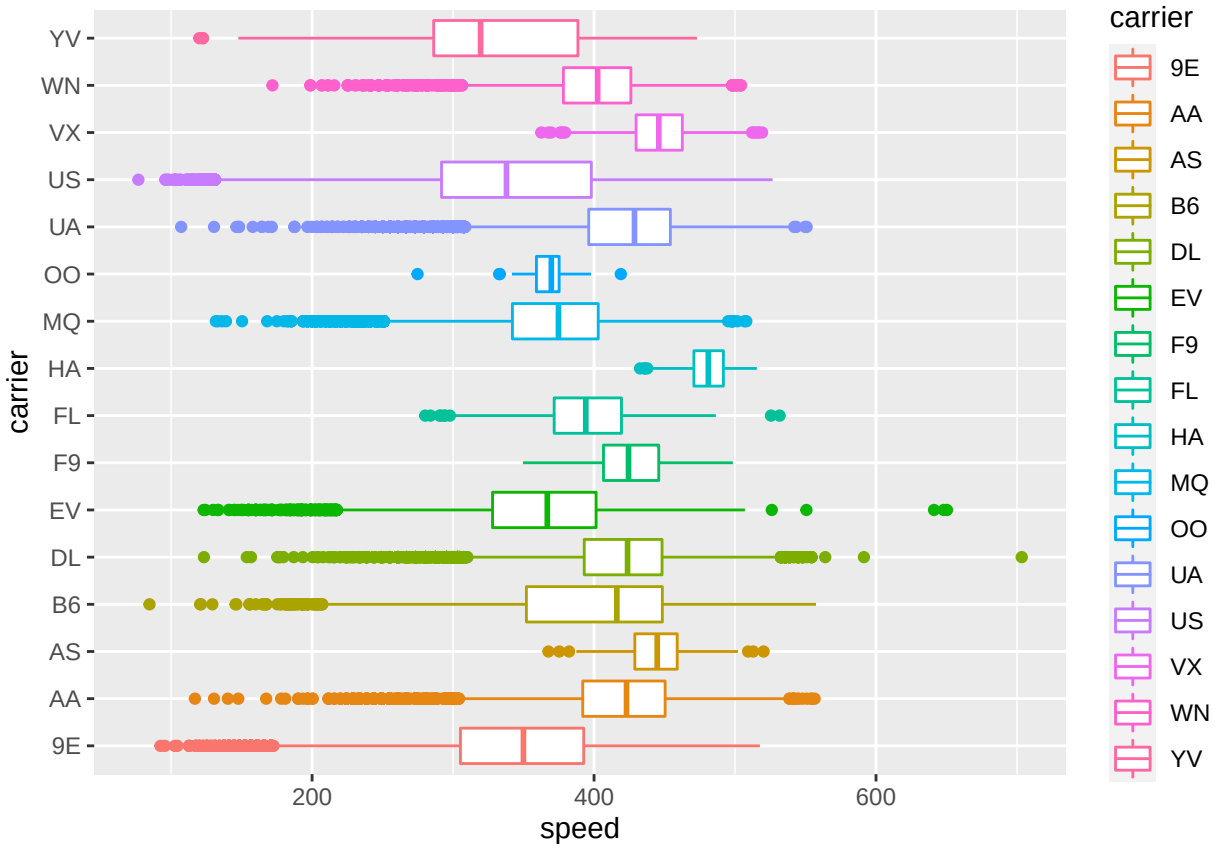
```
head(fast.flights, 3)
```

```
## # A tibble: 3 x 2
## # Groups:   carrier [2]
```

```
##   carrier speed
##   <chr>   <dbl>
## 1 DL       703.
## 2 EV       650.
## 3 EV       648
```

```
ggplot(data = slow.flights, mapping = aes(x = speed, y = carrier, col = carrier)) +
  geom_boxplot()
```

```
## Warning: Removed 9430 rows containing non-finite values (stat_boxplot).
```

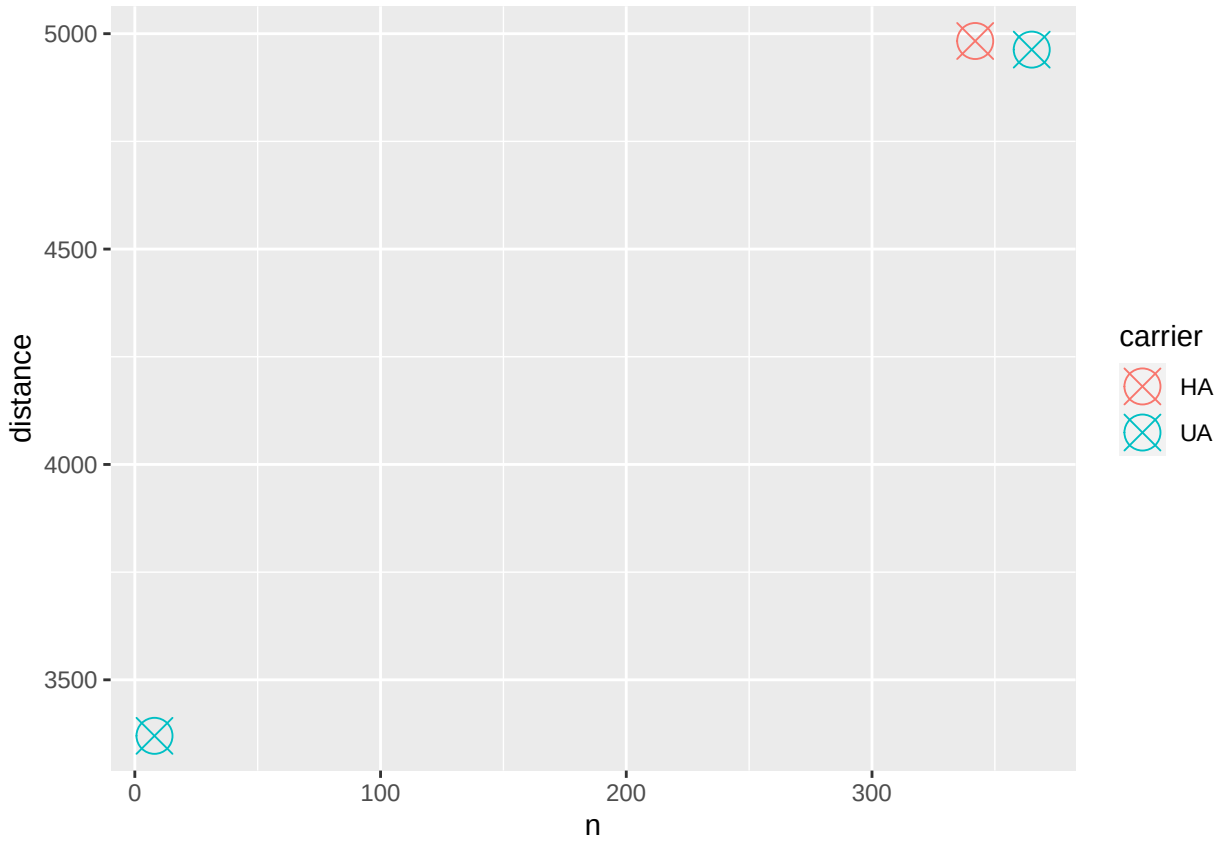


Planteamiento 2:

- El top 3 de los vuelos con mas distancia son de las compañia HA con 4983 Km y 342 vuelos, y la compañias UA con 4963 km con 365 vuelos y 3370 km y 8 vuelos respectivamente.

```
(group_by(flights, carrier, distance) %>%
  summarise(n = n()) %>%
  arrange(desc(distance)))[1:3,] %>%
  ggplot(mapping = aes(x = n, y = distance, type = 21, col = carrier)) +
    geom_point(shape = 13, size = 6)
```

```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```



Pregunta 2

Da una versión equivalente a las pipes siguientes sin usar la función count:

not_cancelled %>% count(dest) not_cancelled %>% count(tailnum, wt = distance)

```
not_cancelled <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))
```

```
# not_cancelled %>% count(dest)
summarise(group_by(not_cancelled, dest), n())
```

```
## # A tibble: 104 x 2
##    dest  `n()`
##    <chr> <int>
##  1 ABQ     254
##  2 ACK     264
##  3 ALB     418
##  4 ANC       8
##  5 ATL   16837
##  6 AUS    2411
##  7 AVL     261
##  8 BDL     412
##  9 BGR     358
## 10 BHM     269
## # ... with 94 more rows
```

```
# not_cancelled %>% count(tailnum, wt = distance)
summarise(group_by(not_cancelled, tailnum), n = sum(distance))
```

```
## # A tibble: 4,037 x 2
##    tailnum      n
##    <chr>    <dbl>
##  1 D942DN    3418
##  2 N0EGMQ  239143
##  3 N10156  109664
##  4 N102UW   25722
##  5 N103US   24619
##  6 N104UW   24616
##  7 N10575  139903
##  8 N105UW   23618
##  9 N107US   21677
## 10 N108UW   32070
## # ... with 4,027 more rows
```

Pregunta 3

Para definir un vuelo cancelado hemos usado la función

(is.na(dep_delay) | is.na(arr_delay))

Intenta dar una definición que sea mejor, ya que la nuestra es un poco subóptima. ¿Cuál es la columna más importante?

```
cancelled <- flights %>%
  filter(is.na(dep_delay) | is.na(arr_delay))
cancelled
```

```
## # A tibble: 9,430 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1     1525           1530        -5     1934           1805
##  2  2013     1     1     1528           1459        29     2002           1647
##  3  2013     1     1     1740           1745        -5     2158           2020
##  4  2013     1     1     1807           1738        29     2251           2103
##  5  2013     1     1     1939           1840        59       29           2151
##  6  2013     1     1     1952           1930        22     2358           2207
##  7  2013     1     1     2016           1930        46       NA           2220
##  8  2013     1     1       NA           1630        NA       NA           1815
##  9  2013     1     1       NA           1935        NA       NA           2240
## 10  2013     1     1       NA           1500        NA       NA           1825
## # ... with 9,420 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Pregunta 4

Investiga si existe algún patrón del número de vuelos que se cancelan cada día. Investiga si la proporción de vuelos cancelados está relacionada con el retraso promedio por día en los vuelos. Investiga si la proporción

de vuelos cancelados está relacionada con el retraso promedio por aeropuerto en los vuelos. ¿Qué compañía aérea sufre los peores retrasos?

Sol

Los días 8, 9 y 10 de cada mes hay mas cancelaciones de vuelos.

```
cancelled_prop = round(dim(cancelled)[1] / dim(not_cancelled)[1]*100,2)
cancelled_prop
```

```
## [1] 2.88
```

```
group_by(cancelled, day) %>%
  summarise(n = n(),
            daily_prop = round(n/sum(day)*100,2),
            mean_dep_delay = mean(dep_delay > 0, na.rm = T)) %>%
  arrange(desc(n))
```

```
## # A tibble: 31 x 4
##      day     n daily_prop mean_dep_delay
##    <int> <int>      <dbl>          <dbl>
##  1     8   963      12.5           0.738
##  2     9   626      11.1           0.545
##  3    10   613      10             0.744
##  4    12   473       8.33          0.763
##  5    23   455       4.35          0.756
##  6    28   379       3.57          0.548
##  7    11   376       9.09          0.424
##  8     7   374      14.3           0.589
##  9    22   360       4.55          0.679
## 10     6   334      16.7           0.526
## # ... with 21 more rows
```

```
# ¿Qué compañía aérea sufre los peores retrasos?

select(flights, carrier, dep_delay) %>%
  arrange(desc(dep_delay))
```

```
## # A tibble: 336,776 x 2
##    carrier dep_delay
##    <chr>       <dbl>
##  1 HA           1301
##  2 MQ           1137
##  3 MQ           1126
##  4 AA           1014
##  5 MQ           1005
##  6 DL            960
##  7 DL            911
##  8 DL            899
##  9 DL            898
## 10 AA            896
## # ... with 336,766 more rows
```