Análise Exploratória dos Aeroportos de Nova Iorque

Adriano Abelaira Paz

05/10/2019

Estudo de Caso

Objetivo: Manipulação e Visualização de dados utilizando os pacotes do **tidyverse**.

Comentários iniciais: Imagine que você trabalha na empresa que administra o aeroporto de Nova Iorque. Baseado no grande número de informações geradas, seu chefe lhe pediu *insights* que possam ajudá-lo, por exemplo, a aumentar o faturamento, diminuir custos e otimizar a operação do aeroporto.

Principais preocupações da direção da empresa:

1. Os atrasos de saída e chegada de aeronaves (soma desses atrasos maior que 90 minutos) geram um custo muito grande para empresa e a diretoria gostaria de entender um pouco mais sobre esse problema;
2. O aeroporto La Guardia é ideal para voos domésticos e de curta duração pois são mais cheios e aumentam consideravelmente a venda nas lojas do aeroporto. Pensando nisso, a diretoria quer proibir os voos com duração maior que três horas;
3. A área de manutenção e verificação das aeronaves deseja analisar a participação de cada fabricante de aeronaves nos aeroportos da cidade de Nova Iorque. Assim, a área precisa da relação entre as informações de cada voo e o fabricante do avião;
4. A área financeira sabe que operar aviões maiores e com mais motores é mais caro e que por isso, esses aviões deveriam ter mais assentos, contudo ela acha que as empresas aéreas não entenderam muito bem isso e que essa relação (total de assentos/ total de motores) não está clara;
5. Uma grande preocupação na área da aviação é a visibilidade do piloto. Sabe-se que os estudos sobre as condições climáticas influenciam diretamente nessa variável e "a diretoria" gostaria de entender melhor essa relação.

O que se tem são informações sobre voos de NYC (por exemplo, EWR, JFK e LGA) em 2013: 336.776 voos no total. Para ajudar a entender o que causa atrasos, por exemplo, também foram disponibilizados outros conjuntos de dados úteis. Você possui as seguintes tabelas de dados:

- flights/voos: todos os voos que partiram de Nova Iorque em 2013
- weather/tempo: dados meteorológicos de hora em hora para cada aeroporto
- planes/avioes: informações de construção sobre cada plano
- airports/aeroportos: nomes e localizações de aeroportos
- airlines/companhias aéreas: tradução entre códigos e nomes de transportadora de duas letras

Descrição das variáveis: Antes de usar as bases importadas faremos uma breve exploração das mesmas.

1. Flights:
year, month, day: data do voo;
dep_time: horário de saída do voo;
sched_dep_time: horário de saída agendado do voo;
dep_delay: atraso na saída em minutos. Valores negativos representam saída antecipada;
arr_time: horário de chegada do voo;
sched_arr_time: horário de chegada agendado do voo;
arr_delay: atraso na chegada em minutos. Valores negativos representam chegadas antecipadas;
carrier: sigla da companhia aérea;
flight: número do voo;
tailnum: número da cauda do avião;
origin: origem do voo;
dest: destino do voo;
air_time: duração do voo em minutos;
distance: distância do voo;
hour, minute: hora e minutos do voo;
time_hour: data e hora agendada do voo no formato POSIXct.

2. Weather:
origin: estação meteorológica;
year, month, day, hour: ano, mês, dia e hora de registro;
temp: temperatura em F;
dewp: ponto de condensação da água em F;
humid: humidade relativa;
wind_dir: direção do vento em graus;
wind_speed: velocidade do vento em mph;
wind_gust: velocidade da rajada de vento em mph;
precip: precipitação em polegadas;
pressure: pressão em milibares;
visib: visibilidade em milhas;
time_hour: data e hora do registro no formato POSIXct.

3. Planes:
tailnum: número da cauda do avião;
year: ano de fabricação do avião;
type: tipo de aeronave;
manufacturer: fabricante;
model: modelo;
engines: número de motores;
seats: número de assentos;
engine: tipo de motor;
speed: média de velocidade voando em mph.

4. Airports:
faa: código FAA do aerooprto;
name: nome do aeroporto;
lat: coordenadas da latitude do aeroporto;
lon: coordenadas da longitude do aeroporto;
alt: altitude, em pés;
tz: deslocamento do fuso horário pela GMT;
dst: horário de verão (A= horário padrão US, U= Unknown, N= no dst);
tzone: fuso horário da IANA, conforme determinado pelo webservice do GeoNames.

5. Airlines:
carrier: sigla da companhia aéreas;

name: nome da companhia aérea.

Observação importante: os arquivos encontram-se em diferentes formatos como: CSV ENG, CSV POR, JSON, rdata, SQL etc. A primeira tarefa é importar e organizar os dados, mas antes o diretório de trabalho onde os arquivos se encontram, deve ser setado.

Lendo os dados de Flights.

```
flight <- read_csv("../data/flights.csv")
flight <- subset(flight, select = -X1); flight
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Lendo os dados de Weather.

```
weather <- base::readRDS("../data/weather.rds"); weather
```

```
## # A tibble: 26,115 x 15
##    origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr>  <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>
##  1 EWR     2013     1     1     1  39.0  26.1  59.4      270      10.4
##  2 EWR     2013     1     1     2  39.0  27.0  61.6      250       8.06
##  3 EWR     2013     1     1     3  39.0  28.0  64.4      240      11.5
##  4 EWR     2013     1     1     4  39.9  28.0  62.2      250      12.7
##  5 EWR     2013     1     1     5  39.0  28.0  64.4      260      12.7
##  6 EWR     2013     1     1     6  37.9  28.0  67.2      240      11.5
##  7 EWR     2013     1     1     7  39.0  28.0  64.4      240      15.0
##  8 EWR     2013     1     1     8  39.9  28.0  62.2      250      10.4
##  9 EWR     2013     1     1     9  39.9  28.0  62.2      260      15.0
## 10 EWR     2013     1     1    10  41    28.0  59.6      260      13.8
## # ... with 26,105 more rows, and 5 more variables: wind_gust <dbl>, precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

Lendo os dados de Planes.

```
plane <- jsonlite::fromJSON("../data/planes.json")
plane <- plane %>% mutate(manufacturer = str_trim(str_remove_all(manufacturer,"INDUSTRIE")))
plane <- plane %>% mutate(manufacturer = str_trim(str_replace_all(manufacturer,"CANADAIR LTD","BOMBARDI
```

```
plane <- plane %>% mutate(manufacturer = str_trim(str_replace_all(manufacturer,"CANADAIR$","BOMBARDIER
plane <- plane %>% mutate(manufacturer =
                          str_trim(str_replace_all(manufacturer,"MCDONNELL DOUGLAS AIRCRAFT CO","MCDO
plane <- plane %>% mutate(manufacturer =
                          str_trim(str_replace_all(manufacturer,"MCDONNELL DOUGLAS CORPORATION","MCDO
plane <- plane %>% mutate(manufacturer = str_trim(str_replace_all(manufacturer,"^DOUGLAS","MCDONNELL DO
(as_tibble(plane))
```

```
## # A tibble: 3,322 x 9
##    tailnum  year type           manufacturer model   engines seats engine  speed
##    <chr>   <int> <chr>          <chr>        <chr>     <int> <int> <chr>   <int>
##  1 N10156   2004 Fixed wing mu... EMBRAER      EMB-14...       2    55 Turbo-...    NA
##  2 N102UW   1998 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  3 N103US   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  4 N104UW   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  5 N10575   2002 Fixed wing mu... EMBRAER      EMB-14...       2    55 Turbo-...    NA
##  6 N105UW   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  7 N107US   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  8 N108UW   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
##  9 N109UW   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
## 10 N110UW   1999 Fixed wing mu... AIRBUS       A320-2...       2   182 Turbo-...    NA
## # ... with 3,312 more rows
```

Lendo dados de Airports.

```
airport <- read_delim("../data/airports.txt", delim = " ", col_names = FALSE, skip = 1)
airport <- airport[,-1]
colnames(airport) <- colnames(read_delim('../data/airports.txt', delim = " ", n_max = 0))
airport
```

```
## # A tibble: 1,458 x 8
##    faa   name                      lat    lon   alt    tz dst   tzone
##    <chr> <chr>                   <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
##  1 04G   Lansdowne Airport        41.1  -80.6  1044    -5 A     America/New_Yo...
##  2 06A   Moton Field Municipal A...  32.5  -85.7   264    -6 A     America/Chicago
##  3 06C   Schaumburg Regional      42.0  -88.1   801    -6 A     America/Chicago
##  4 06N   Randall Airport          41.4  -74.4   523    -5 A     America/New_Yo...
##  5 09J   Jekyll Island Airport    31.1  -81.4    11    -5 A     America/New_Yo...
##  6 0A9   Elizabethton Municipal ...  36.4  -82.2  1593    -5 A     America/New_Yo...
##  7 0G6   Williams County Airport  41.5  -84.5   730    -5 A     America/New_Yo...
##  8 0G7   Finger Lakes Regional A...  42.9  -76.8   492    -5 A     America/New_Yo...
##  9 0P2   Shoestring Aviation Air...  39.8  -76.6  1000    -5 U     America/New_Yo...
## 10 0S9   Jefferson County Intl    48.1 -123.    108    -8 A     America/Los_An...
## # ... with 1,448 more rows
```

Lendo os dados de Airlines.

```
airline1 <- read_delim("../data/airlines_1.csv", delim = ";")
airline2 <- read_delim(file = "../data/airlines_2.csv", delim = ";")
airline3 <- read_delim("../data/airlines_2.csv", delim = ";")
airline <- rbind(airline1, airline2, airline3)
airline <- subset(airline, select = -X1) %>% distinct()
airline
```

```
## # A tibble: 10 x 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
```

Primeira preocupação:

> Os atrasos de saída e chegada de aeronaves (soma desses atrasos maior que 90 minutos) geram um custo muito grande para empresa e a diretoria gostaria de entender um pouco mais sobre esse problema.

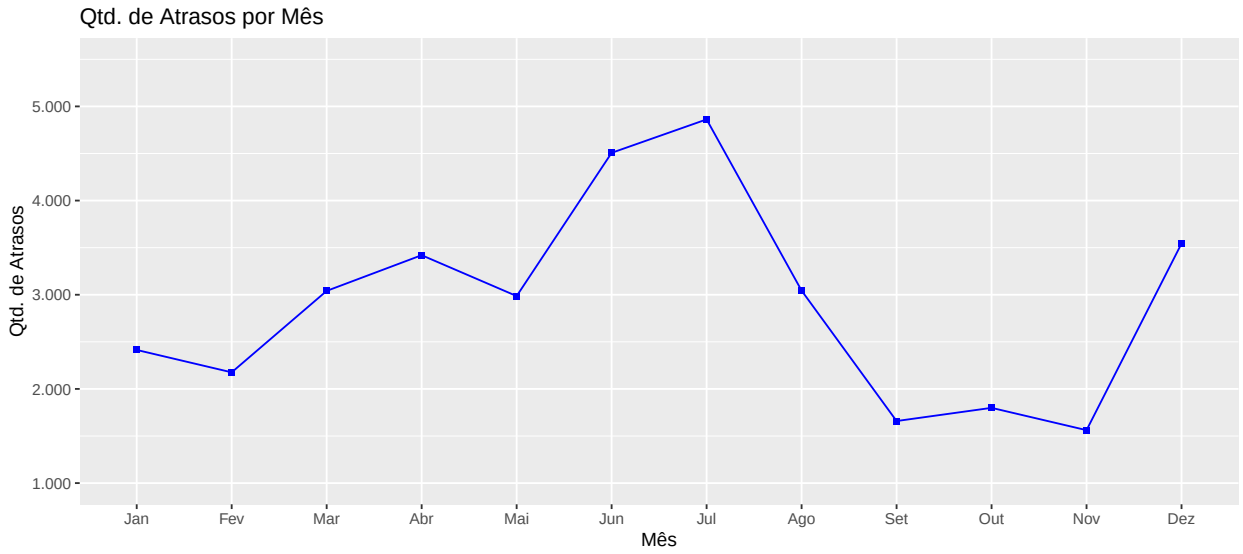
Obtendo um *Data Set* com os minutos de atrasos maiores que 90 minutos para todos os voos.

```
flight_delay <- flight %>% mutate(atraso_total = dep_delay + arr_delay)
flight_delay <-  flight_delay %>% filter(atraso_total > 90) %>%
  select(year,month,day,carrier,tailnum,flight,origin,dest,atraso_total)
flight_delay
```

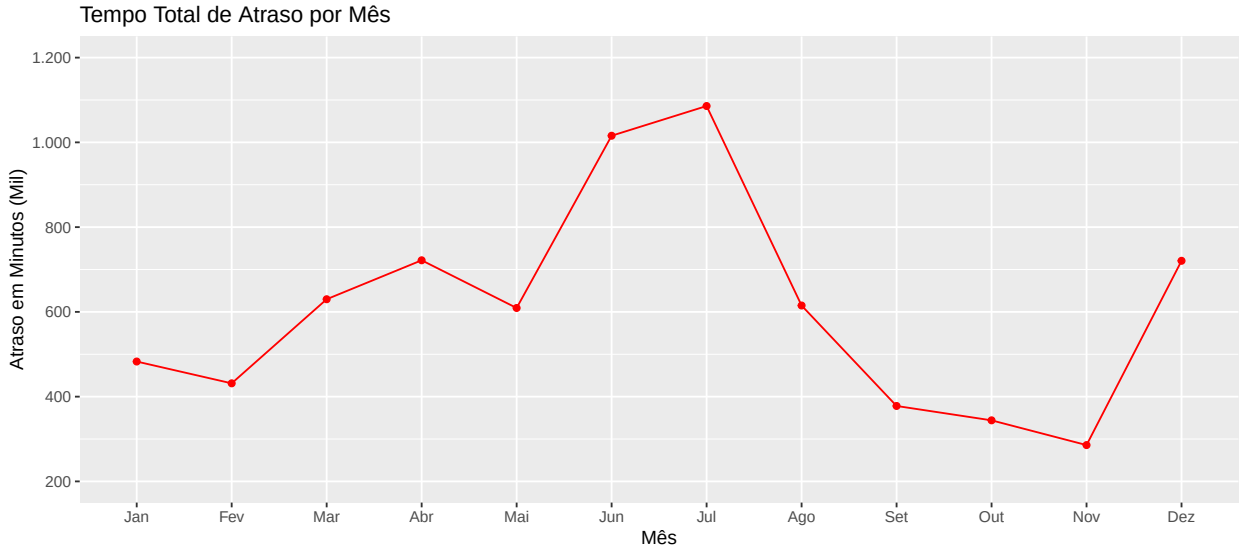
```
## # A tibble: 35,014 x 9
##     year month   day carrier tailnum flight origin dest  atraso_total
##    <dbl> <dbl> <dbl> <chr>   <chr>    <dbl> <chr>  <chr>        <dbl>
##  1  2013     1     1 MQ      N531MQ    4576 LGA    CLT            238
##  2  2013     1     1 AA      N3GVAA     443 JFK    MIA            122
##  3  2013     1     1 MQ      N942MQ    3944 JFK    BWI           1704
##  4  2013     1     1 MQ      N532MQ    4655 LGA    BNA             93
##  5  2013     1     1 UA      N534UA     856 EWR    BOS            267
##  6  2013     1     1 UA      N76502    1086 LGA    IAH            279
##  7  2013     1     1 EV      N16561    4495 EWR    SAV            174
##  8  2013     1     1 MQ      N518MQ    4601 LGA    BNA            136
##  9  2013     1     1 MQ      N542MQ    4646 LGA    MSP            164
## 10  2013     1     1 EV      N19966    4640 EWR    DAY             93
## # ... with 35,004 more rows
```

Investigando as relações dos minutos de atraso com a data do voo.

```
flight_delay %>% group_by(month) %>% summarise(qtd = n()) %>%
  ggplot(aes(x = month, y = qtd)) +
  geom_point(shape = 15, colour = 'blue') + geom_line(colour = 'blue') +
  scale_y_continuous(limits = c(1000,5500),
                     breaks = c(1000,2000,3000,4000,5000),
                     labels = c('1.000','2.000','3.000','4.000','5.000')) +
  scale_x_discrete(limits = c('1','2','3','4','5','6','7','8','9','10','11','12'),
                   labels = c('Jan','Fev','Mar','Abr','Mai','Jun','Jul','Ago','Set','Out','Nov','Dez')) +
  labs(x= "Mês", y = "Qtd. de Atrasos", title = "Qtd. de Atrasos por Mês")
```




```
flight_delay %>% group_by(month) %>% summarise(total = sum(atraso_total)) %>%
  ggplot(aes(x = month, y = total)) +
  geom_point(colour = 'red') + geom_line(colour = 'red') +
  scale_y_continuous(limits = c(200e3,1200e3),
                     breaks = c(200e3,400e3,600e3,800e3,1000e3,1200e3),
                     labels = c('200','400','600','800','1.000','1.200')) +
  scale_x_discrete(limits = c('1','2','3','4','5','6','7','8','9','10','11','12'),
                   labels = c('Jan','Fev','Mar','Abr','Mai','Jun','Jul','Ago','Set','Out','Nov','Dez')) +
  labs(x = "Mês", y = "Atraso em Minutos (Mil)", title = "Tempo Total de Atraso por Mês")
```



Verificando se existe alguma influência do tempo no mês de Julho que explique essa valor alto.

Obtendo as médias das métricas de tempo para todos os meses, com exceção do mês de Julho.

```
weather %>% filter(month != 7) %>%
    summarise(temp_mean = mean(temp, na.rm = TRUE),
              dewp_mean = mean(dewp, na.rm = TRUE),
```

```
            humid_mean = mean(humid, na.rm = TRUE),
            wind_dir_mean = mean(wind_dir, na.rm = TRUE),
            wind_speed_mean = mean(wind_speed, na.rm = TRUE),
            wind_gust_mean = mean(wind_gust, na.rm = TRUE),
            precip_mean = mean(precip, na.rm = TRUE),
            pressure_mean = mean(pressure, na.rm = TRUE),
            visib_mean = mean(visib, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   temp_mean dewp_mean humid_mean wind_dir_mean wind_speed_mean wind_gust_mean
##       <dbl>     <dbl>      <dbl>         <dbl>           <dbl>          <dbl>
## 1      52.9      39.1       62.1          201.            10.6           25.7
## # ... with 3 more variables: precip_mean <dbl>, pressure_mean <dbl>,
## #   visib_mean <dbl>
```

Comparando as médias das métricas do tempo para todos os meses, com exceção do mês de Julho, com as médias das métricas para o mês de Julho.

```
weather %>% filter(month == 7) %>%
  summarise(temp_mean = mean(temp),
            dewp_mean = mean(dewp),
            humid_mean = mean(humid),
            wind_dir_mean = mean(wind_dir, na.rm = TRUE),
            wind_speed_mean = mean(wind_speed, na.rm = TRUE),
            wind_gust_mean = mean(wind_gust, na.rm = TRUE),
            precip_mean = mean(precip),
            pressure_mean = mean(pressure, na.rm = TRUE),
            visib_mean = mean(visib))
```

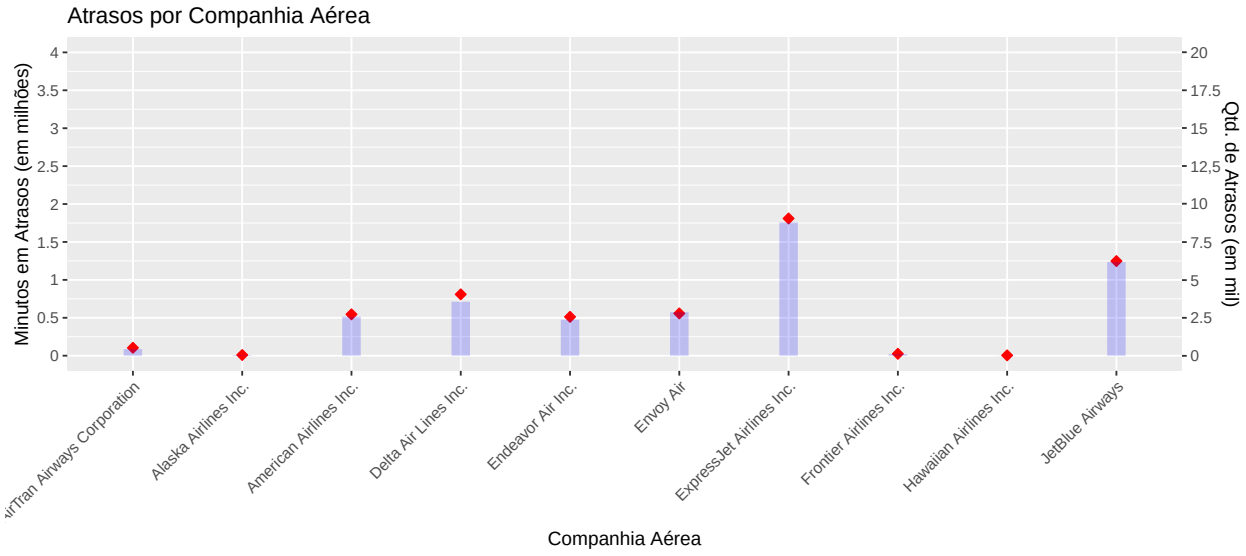
```
## # A tibble: 1 x 9
##   temp_mean dewp_mean humid_mean wind_dir_mean wind_speed_mean wind_gust_mean
##       <dbl>     <dbl>      <dbl>         <dbl>           <dbl>          <dbl>
## 1      80.1      67.0       66.9          191.            9.58           21.4
## # ... with 3 more variables: precip_mean <dbl>, pressure_mean <dbl>,
## #   visib_mean <dbl>
```

Analisando o resultado, chegou-se ao entendimento que nos meses de Junho e Julho existe uma grande concentração de quantidade de atrasos. Essa concentração pela quantidade também é validada pelo aspecto do total de minutos em atraso. Não foi identificada quaisquer influência do tempo que justifique essa alta nos meses de Junho e Julho.
Uma explicação para esse fenômeno seria as férias de verão no hemisfério norte que causa um incremento no número de viagens e aumenta a possibilidade de atrasos.

Investigando se existe uma relação dos minutos de atraso com as companhias aéreas.

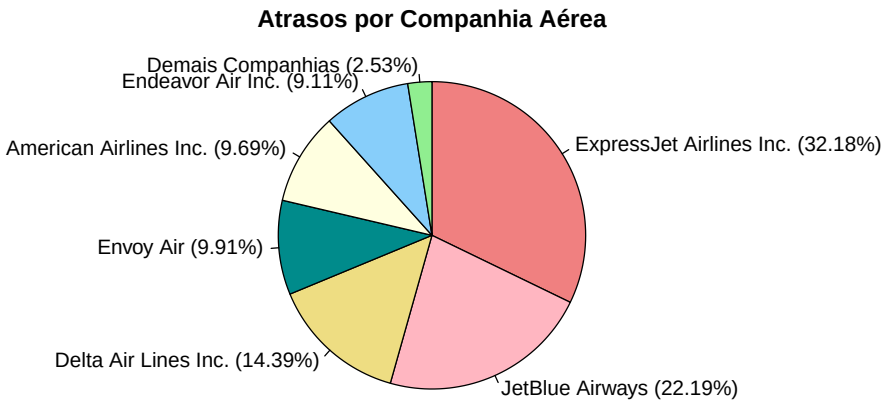
```
flight_delay <- flight_delay %>%
  inner_join(airline, by = c('carrier' = 'carrier')) %>% mutate(airline = name) %>%
  select(year,month,day,carrier,airline,tailnum,flight,origin,dest,atraso_total)

flight_airline <- group_by(flight_delay, airline) %>%
  summarise(total_atraso = sum(atraso_total), atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  arrange(desc(total_atraso)) %>%
```

```
  mutate(perc_atraso = total_atraso / sum(total_atraso))

ggplot(flight_airline, aes(x = as.factor(airline))) +
  geom_point(aes(y = total_atraso), colour = 'red', shape = 23, fill = 'red', size = 2) +
  geom_segment(aes(x = as.factor(airline), xend = as.factor(airline), y = 0, yend = qtd_atraso*200),
               colour = 'blue', size = 5, alpha = 0.2) +
  scale_y_continuous(limits = c(0,4000000),
                     breaks = c(0,0.5e06,1e06,1.5e06,2e06,2.5e06,3e06,3.5e06,4e06),
                     labels = c("0","0.5","1","1.5","2","2.5","3","3.5","4"),
                     sec.axis = sec_axis(~./200, name = "Qtd. de Atrasos (em mil)",
                                         breaks = c(0,2500,5000,7500,10000,12500,15000,17500,20000),
                                         labels = c('0','2.5','5','7.5','10','12,5','15','17.5','20'))) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = c(0.1,0.1)) +
  labs(x = "Companhia Aérea", y = "Minutos em Atrasos (em milhões)",
       title = "Atrasos por Companhia Aérea")
```



O gráfico acima plota a relação das companhias aéreas com atraso em minutos, elemento gráfico em azul com eixo na esquerda, e o número de atrasos, elemento gráfico em vermelho com eixo na direita.
Percebe-se um grande destaque para a companhia aérea ExpressJet Airlines Inc, seguida pela JetBlue Airways. Os atrasos dessas duas companhias se destacam das demais, tanto em minutos totais, quanto em quantidade de atrasos.

```
flight_airline_agr <- bind_rows(flight_airline %>% filter(perc_atraso >= 0.05),
                                flight_airline %>% filter(perc_atraso < 0.05) %>%
                                summarise(airline = "Demais Companhias", total_atraso = sum(total_atraso
                                          atraso_medio = mean(atraso_medio),
                                          qtd_atraso = sum(qtd_atraso), perc_atraso = sum(perc_atraso))

pie(flight_airline_agr$perc_atraso,
    labels = paste(flight_airline_agr$airline," (",round(flight_airline_agr$perc_atraso*100,2),"%)",sep
    main = "Atrasos por Companhia Aérea", clockwise = TRUE, radius = 1,
    col = c('lightcoral','lightpink','lightgoldenrod','cyan4','lightyellow','lightskyblue','lightgreen'
```



Essa diferença nos atrasos das companhias ExpressJet Airlines Inc e JetBlue Airways, em relação as demais, fica evidente no gráfico acima.
As duas companhias representam mais de 50% dos minutos totais de atrasos maior que 90 minutos em comparação com as demais.

Investigando com mais detalhes a causa dos atrasos das companhias ExpressJet Airlines Inc e JetBlue Airways.

```
flight_airline_ejb6 <- flight_delay %>% filter(carrier %in% c('B6','EV'))

flight_airline_ejb6 <- flight_airline_ejb6 %>% left_join(plane, by = c('tailnum','tailnum')) %>%
  select(year.x,month,day,carrier,airline,tailnum,model,year.y,flight,origin,dest,atraso_total)
colnames(flight_airline_ejb6) <-
  c('year_flight','month','day','carrier','airline','tailnum','model','year_model','flight','origin','d

ggplot(flight_airline_ejb6, aes(x = as.factor(model), y = atraso_total)) +
  geom_jitter(aes(colour = as.factor(flight_airline_ejb6$year_model))) +
  scale_colour_viridis_d(na.value = "gray75") +
  scale_y_continuous(limits = c(0,1200),
                     breaks = c(0,100,200,300,400,500,600,700,800,900,1000,1100,1200)) +
  scale_x_discrete(na.translate = FALSE) +
  labs(x = "Modelo do Avião", y = "Minutos de Atraso", colour = "Ano") +
  theme(legend.position = 'bottom', axis.text.x = element_text(angle = 45, hjust = 1))
```

O gráfico acima fornece uma indicação que os modelos **EMB-145-LR**, **EMB-145XR** e **ERJ 190-100 IGW** da Embraer e o modelo **A320-232** da Airbus sejam os principais responsáveis pelos atrasos das companhias ExpressJet Airlines Inc e JetBlue Airways.

Obtendo uma relação de medidas de posição e dispersão agrupados por modelo para a companhia ExpressJet Airines Inc.

```
## [1] "Quadro de Medidas de Posição e Dispersão Geral"
```

```
(t <- flight_airline_ejb6 %>% filter(carrier == 'EV') %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(total_atraso,qtd_atraso,atraso_medio,desvioPad))
```

```
## # A tibble: 1 x 4
##   total_atraso qtd_atraso atraso_medio desvioPad
##          <dbl>      <int>        <dbl>     <dbl>
## 1      1809589       8784         206.      116.
```

```
## [1] "Quadro de Medidas de Posição e Dispersão por Modelo de avião"
```

```
(x <- flight_airline_ejb6 %>% filter(carrier == 'EV') %>%
    group_by(model) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(model,total_atraso,qtd_atraso,atraso_medio,desvioPad))
```

```
## # A tibble: 6 x 5
##   model       total_atraso qtd_atraso atraso_medio desvioPad
##   <chr>              <dbl>      <int>        <dbl>     <dbl>
## 1 CL-600-2B19        90973        359         253.      153.
## 2 CL-600-2C10       255674       1211         211.      121.
## 3 CL-600-2D24        28859        133         217.      120.
## 4 EMB-145             8072         42         192.       70.6
## 5 EMB-145LR         966665       4785         202.      112.
## 6 EMB-145XR         459346       2254         204.      112.
```

```
## [1] "Percentual de atraso em minutos para modelo EMB-145LR: 53.42%"
```

Nos quadros acima, nota-se que o modelo **EMB-145LR** da Embraer é responável por cerca de 50% do total de atraso em minutos.

Retirando o modelo **EMB-145-LR** da Embraer da amostra, verificamos uma melhora sensível das métricas, total_atraso, que representa o total de atraso em minutos, e qtd_atraso, que representa o quantidade de atraso.

```
## [1] "Quadro de Medidas de Posição e Dispersão SEM o modelo EMB-145LR"
```

```
flight_airline_ejb6 %>% filter(carrier == 'EV' & model != 'EMB-145LR') %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(total_atraso,qtd_atraso,atraso_medio,desvioPad)
```

```
## # A tibble: 1 x 4
##   total_atraso qtd_atraso atraso_medio desvioPad
##          <dbl>      <int>        <dbl>     <dbl>
## 1       842924       3999         211.      120.
```

Obtendo uma relação de medidas de posição e dispersão agrupados por modelo para a companhia JetBlue Airways.

```
## [1] "Quadro de Medidas de Posição e Dispersão Geral"
```

```
(t <- flight_airline_ejb6 %>% filter(carrier == 'B6') %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(total_atraso,qtd_atraso,atraso_medio,desvioPad))
```

```
## # A tibble: 1 x 4
##   total_atraso qtd_atraso atraso_medio desvioPad
##          <dbl>      <int>        <dbl>     <dbl>
## 1      1247691       6171         202.      113.
```

```
## [1] "Quadro de Medidas de Posição e Dispersão por Modelo de avião"
```

```
(x <- flight_airline_ejb6 %>% filter(carrier == 'B6') %>%
    group_by(model) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(model,total_atraso,qtd_atraso,atraso_medio,desvioPad))
```

```
## # A tibble: 7 x 5
##   model           total_atraso qtd_atraso atraso_medio desvioPad
##   <chr>                  <dbl>      <int>        <dbl>     <dbl>
## 1 A320-232              736401       3664         201.     111.
## 2 A321-231                2026         14         145.      31.8
## 3 ERJ 190-100 IGW       477575       2326         205.     116.
## 4 R66                     5854         32         183.      89.4
## 5 SR22                    4824         23         210.     129.
## 6 ZODIAC 601HDS           5482         30         183.      82.0
## 7 <NA>                   15529         82         189.     107.
```

```
## [1] "Percentual de atraso em minutos para o modelo A320-232: 59.02%"
```

Nos quadros acima, nota-se que o modelo **A320-232** da Airbus é responável por cerca de 60% do total de atraso em minutos.

Retirando o modelo **A320-232** da Airbus da amostra, verificamos uma melhora sensível das métricas, total_atraso, que representa o total de atraso em minutos, e qtd_atraso, que representa o quantidade de atraso.

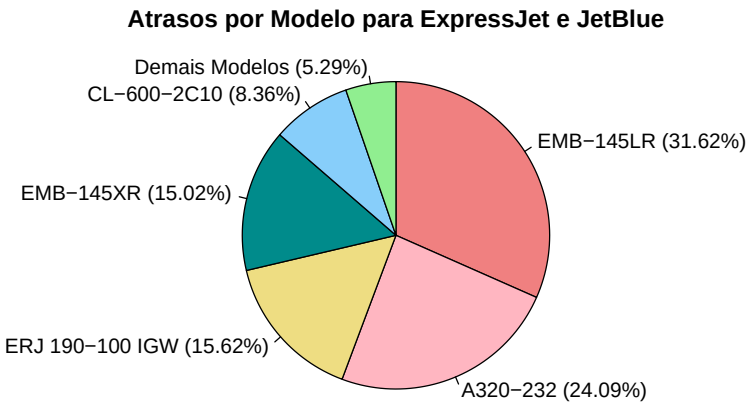
```
## [1] "Quadro de Medidas de Posição e Dispersão SEM o modelo A320-232"
```

```
flight_airline_ejb6 %>% filter(carrier == 'B6' & model != 'A320-232') %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(total_atraso,qtd_atraso,atraso_medio,desvioPad)
```

```
## # A tibble: 1 x 4
##   total_atraso qtd_atraso atraso_medio desvioPad
##          <dbl>      <int>        <dbl>     <dbl>
## 1       495761       2425         204.      115.
```

O gráfico abaixo evidencia a grande contribuição do modelo EMB-145LR e A320-232 no total de atrasos das companhias ExpressJet Airlines Inc e JetBlue Airways respectivamente.

```
flight_airline_ejb6_plane <- flight_airline_ejb6 %>% group_by(model) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  mutate(perc_atraso = total_atraso / sum(total_atraso)) %>%
  arrange(desc(total_atraso)) %>%
  select(model,total_atraso,qtd_atraso,perc_atraso)

flight_airline_ejb6_plane_agr <- bind_rows(flight_airline_ejb6_plane %>% filter(perc_atraso >= 0.05),
  flight_airline_ejb6_plane %>% filter(perc_atraso < 0.05) %>%
  summarise(model = "Demais Modelos", total_atraso = sum(total_atraso),
            qtd_atraso = sum(qtd_atraso), perc_atraso = sum(perc_atraso)))

pie(flight_airline_ejb6_plane_agr$perc_atraso,
    labels = paste(flight_airline_ejb6_plane_agr$model," (",
                   round(flight_airline_ejb6_plane_agr$perc_atraso*100,2),"%)",sep = ""),
    main = "Atrasos por Modelo para ExpressJet e JetBlue", clockwise = TRUE, radius = 1,
    col = c('lightcoral','lightpink','lightgoldenrod','cyan4','lightskyblue','lightgreen'))
```



Investigando mais profundamente a relação de atrasos das companhias com os aeroportos de origem e destino.

O quadro abaixo mostra que não há diferença significante entre a quantidade de voos por aeroporto de origem. A métrica perc_qtd, que representa o percentual da quantidade de atrasos dos voos, permanece estável para os três aeroportos.

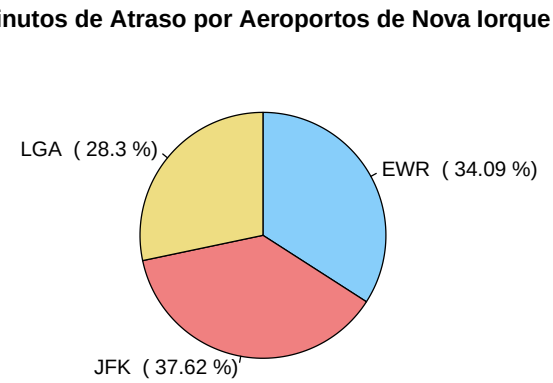
```
## [1] "Quadro com a quantidade total de voos por aeroporto de origem"
```

```
flight %>% group_by(origin) %>%
  summarise(qtd_flight = n()) %>%
  mutate(perc_qtd = (qtd_flight / sum(qtd_flight))*100) %>%
  select(origin,qtd_flight,perc_qtd)
```

```
## # A tibble: 3 x 3
##   origin qtd_flight perc_qtd
##   <chr>       <int>    <dbl>
## 1 EWR        120835     35.9
## 2 JFK        111279     33.0
## 3 LGA        104662     31.1
```

Analisando a distribuição dos minutos de atraso por aeroporto de origem, nota-se que a contribuição de cada aeroporto não se alterou significativamente. Os três aeroportos possuem representatividade semelhanante.

```
flight_delay_airport_origin <- flight_delay %>% group_by(origin) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  mutate(perc_atraso = total_atraso / sum(total_atraso)) %>%
  select(origin,total_atraso,qtd_atraso,perc_atraso,atraso_medio,desvioPad)

pie(flight_delay_airport_origin$perc_atraso,
    labels = paste(flight_delay_airport_origin$origin, " (",
                   round(flight_delay_airport_origin$perc_atraso*100,2),"%)"),
    main = "Minutos de Atraso por Aeroportos de Nova Iorque",
    clockwise = TRUE,
    col = c('lightskyblue','lightcoral','lightgoldenrod'))
```



Filtrando os dados de atraso do aeroporto de Newark Liberty Intl para considerar somente os atrasos das companhias ExpressJet Airlines Inc e JetBlue Airways, Verifica-se que o percentual de atraso que essas companhia possuem, representa cerca de 86% do total de minutos de atrasos para esse aeroporto.

```
## [1] "Atraso (em Minutos) das Companhias ExpressJet e JetBlue no Aeroporto Newark Liberty Intl: 1,650
```

```
## [1] "Percentual de Atraso das Companhias ExpressJet e JetBlue no Aeroporto Newark Liberty Intl: 86.1
```

Avaliando para os demais aeroportos de Nova Iorque, JFK (John F Kennedy Intl) e LGA (La Guardia), constatou-se que as companhias ExpressJet Airlines Inc e JetBlue Airways representam cerca de 50% e 24% dos atrasos em minutos respectivamente.

```
## [1] "Atraso (em Minutos) das Companhias ExpressJet e JetBlue no Aeroporto John F Kennedy Intl: 947,1
```

```
## [1] "Percentual de Atraso das Companhias ExpressJet e JetBlue no Aeroporto John F Kennedy Intl: 49.4
```

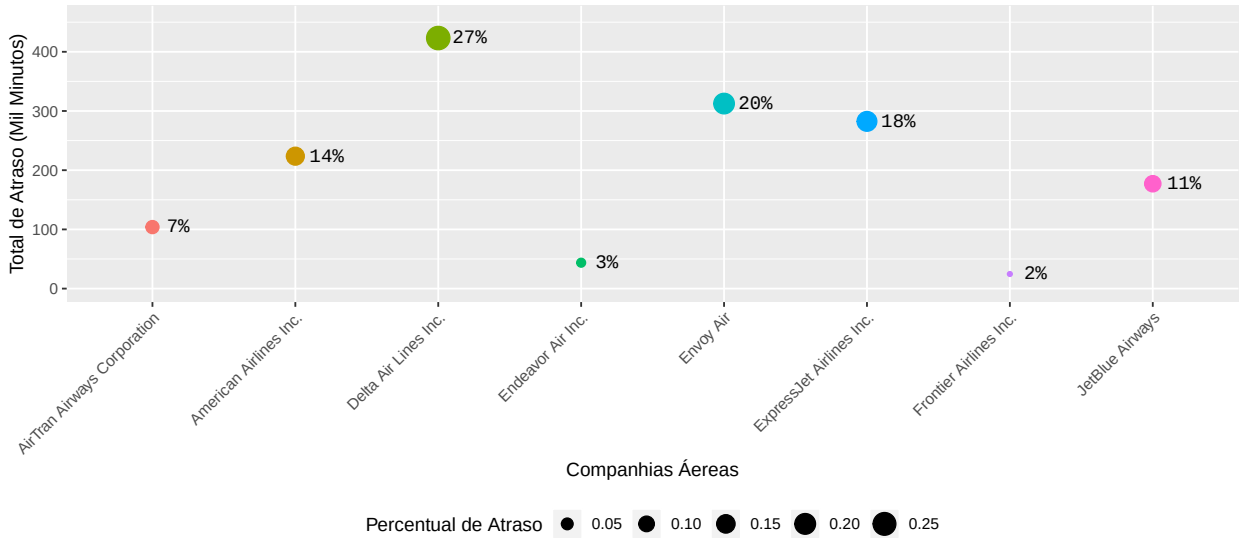
```
## [1] "Atraso (em Minutos) das Companhias ExpressJet e JetBlue no Aeroporto La Guardia: 459,594"
```

```
## [1] "Percentual de Atraso das Companhias ExpressJet e JetBlue no Aeroporto La Guardia: 23.98%"
```

Investigando um pouco mais sobre o aeroporto de origem LGA (La Guardia), percebeu-se que a companhia que se destacou no total de atrasos foi a Delta Air Lines Inc, correspondendo a cerca de 27% dos atrasos nesse aeroporto. Sendo acompanhado pelas companhias Envoy Air e ExpressJet Airlines, com 20% e 18% do total de atrasos para o aeroporto.

```
flight_delay_airport_lga <- flight_delay %>% filter(origin == 'LGA') %>% group_by(airline) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  mutate(perc_atraso = total_atraso / sum(total_atraso)) %>%
  arrange(desc(perc_atraso)) %>%
  select(airline,total_atraso,qtd_atraso,perc_atraso,atraso_medio,desvioPad)

ggplot(flight_delay_airport_lga, aes(x = airline, y = total_atraso)) +
```

```
  geom_point(aes(size = flight_delay_airport_lga$perc_atraso,
                 color = as.factor(flight_delay_airport_lga$airline))) +
  geom_text(label = paste(round(flight_delay_airport_lga$perc_atraso*100,0),"%",sep = ""),
            family = 'mono', hjust = 0, nudge_x = 0.1) +
  scale_y_continuous(limits = c(0,4.55e5),
                     breaks = c(0,1e5,2e5,3e5,4e5),
                     labels = c('0','100','200','300','400')) +
  guides(color = 'none') +
  labs(x = "Companhias Áereas", y = "Total de Atraso (Mil Minutos)", size = "Percentual de Atraso") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "bottom")
```




```
paste('Atraso (em Minutos) da Delta Air Lines Inc no Aeroporto ',
      (airport %>% filter(faa == 'EWR'))$name,": ",
      (flight_delay %>% filter(origin == 'EWR' & carrier %in% c('DL')) %>%
         summarise(total = sum(atraso_total)))$total %>% format(big.mark = ','), sep = '')
```

```
## [1] "Atraso (em Minutos) da Delta Air Lines Inc no Aeroporto Newark Liberty Intl: 100,735"
```

```
paste('Atraso (em Minutos) da Delta Air Lines Inc no Aeroporto ',
      (airport %>% filter(faa == 'JFK'))$name,": ",
      (flight_delay %>% filter(origin == 'JFK' & carrier %in% c('DL')) %>%
         summarise(total = sum(atraso_total)))$total %>% format(big.mark = ','), sep = '')
```

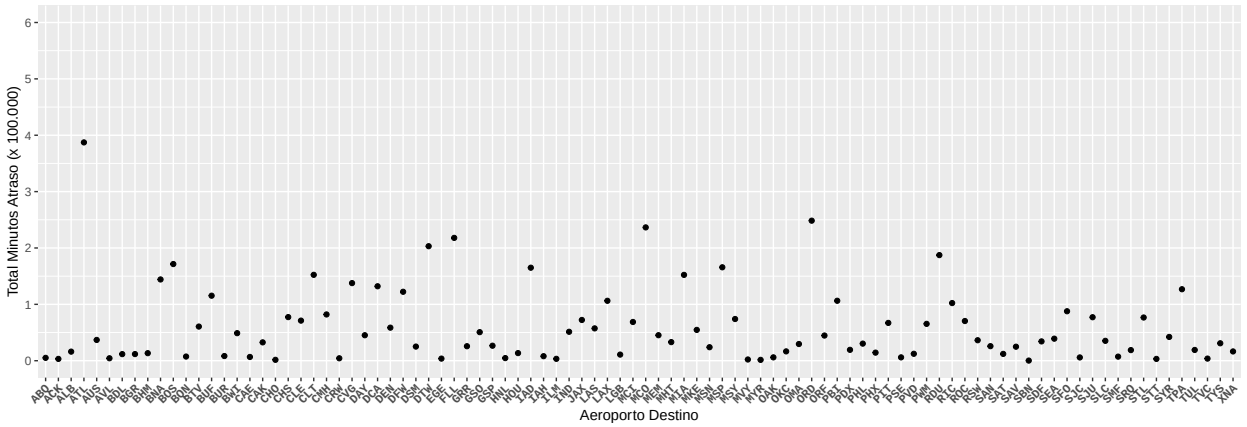
```
## [1] "Atraso (em Minutos) da Delta Air Lines Inc no Aeroporto John F Kennedy Intl: 285,078"
```

Apesar da Delta Air Lines Inc possuir cerca de 400 mil minutos de atraso no aeroporto de La Guardia, esse valor corresponde a menos da metade do total de atraso que as companhias ExpressJet Airlines Inc e JetBlue Airways tiveram no aeroporto John F Kennedy Intl e menos que um terço que elas tiveram no aeroporto de Newark Liberty Intl. Além disso, a diferença entre os totais de atraso, em minutos, da Delta Air Lines Inc em relação as demais companhias no aeroporto de La Guardia não é tão acentuado como nos demais aeroportos de Nova Iorque para as companhias ExpressJet Airlines Inc e JetBlue Airways, acrescendo o fato de que o total de atraso da companhia Delta Air Lines Inc é próximo do total dessas duas companhias no aeroporto de La Guardia. Acrescentando que o total de atraso nos demais aeroportos de Nova Iorque, John

F Kennedy Intl e Newark Liberty Intl, são muito inferiores comparados com a ExpressJet Airlines e JetBlue Airways nesses aeroportos. Posto isso, a companhia Delta Air Lines Inc não foi incluída na conclusão.

Investigando os atrasos para os aeroportos de destino, constatou-se que o aeroporto ATL (Hartsfield Jackson Atlanta Intl) se destaca dos demais aeroportos.

```
flight_delay %>% group_by(dest) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  select(dest,total_atraso,qtd_atraso,atraso_medio,desvioPad) %>%
  ggplot(aes(x = as.factor(dest), y = total_atraso)) +
  geom_point() +
  labs(x = "Aeroporto Destino", y = "Total Minutos Atraso (x 100.000)") +
  scale_y_continuous(limits = c(0,6e5),
                     breaks = c(0,1e5,2e5,3e5,4e5,5e5,6e5),
                     labels = c('0','1','2','3','4','5','6')) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, family = 'mono', face = 'bold'))
```



Aprofundando a análise nos aeroportos de destino, agrupou-se os dados por companhia aérea. Notou-se um destaque para a companhia a Delta Air Lines Inc. Como o total de minutos de atraso não é relavante comparado com o da companhia ExpressJet Airlines Inc mais a JetBlue Airways, essa análise não foi incluída na conclusão.

```
flight_delay %>% filter(dest == 'ATL') %>% group_by(airline) %>%
  summarise(desvioPad = sd(atraso_total), total_atraso = sum(atraso_total),
            atraso_medio = mean(atraso_total), qtd_atraso = n()) %>%
  mutate(perc_atraso = total_atraso / sum(total_atraso)) %>%
  select(airline,total_atraso,perc_atraso,qtd_atraso,atraso_medio,desvioPad)
```

```
## # A tibble: 5 x 6
##   airline             total_atraso perc_atraso qtd_atraso atraso_medio desvioPad
##   <chr>                      <dbl>       <dbl>      <int>        <dbl>     <dbl>
## 1 AirTran Airways Co...      71239    0.184           301         237.      174.
## 2 Delta Air Lines In...     207366    0.535           861         241.      185.
## 3 Endeavor Air Inc.            91    0.000235          1          91       NA
## 4 Envoy Air                 46082    0.119           250         184.      93.6
## 5 ExpressJet Airline...     62526    0.161           303         206.      114.
```

Conclusão:

Existe uma grande concentração de atrasos nos meses de Junho e Julho, sendo que essa concentração se deve, principalmente, ao número de voos que aumenta abruptamente nas férias de verão no hemisfério norte.
A companhia ExpressJet Airlines Inc. e JetBlue Airways são responsáveis por mais da metade dos atrasos ocorridos nos aeroportos administrados pela empresa. Dentre os atrasos dessas companhias, o modelo EMB-145LR da Embraer, que pertence a ExpressJet Airlines Inc, e o modelo A320-232 da Airbus, que pertence a JetBlue Airways, são responsáveis por mais da metade deles. Não há grande destaques nos atrasos entre os aeroportos da empresa. Cada um contribuiu com cerca de 1/3 de atrasos no total. No aeroporto de Newark Liberty Intl as companhias ExpressJet Airlines Inc e JetBlue Airways foram responsáveis por mais de 85% deles. No aeroporto John F Kennedy Intl, as duas companhais foram responsáveis por cerca de 50% e no La Guardia por cerca de 25% dos atrasos em minutos.

```
## [1] "Quadro com os atrasos para a companhia ExpressJet Airlines Inc"
```

```
## # A tibble: 6 x 5
##   model       total_atraso qtd_atraso atraso_medio desvioPad
##   <chr>              <dbl>      <int>        <dbl>     <dbl>
## 1 CL-600-2B19        90973        359         253.     153.
## 2 CL-600-2C10       255674       1211         211.     121.
## 3 CL-600-2D24        28859        133         217.     120.
## 4 EMB-145             8072         42         192.      70.6
## 5 EMB-145LR         966665       4785         202.     112.
## 6 EMB-145XR         459346       2254         204.     112.
```

```
## [1] "Quantidade de voos para o modelo EMB-145XR: "
```

```
## [1] 14051
```

```
## [1] "Quantidade de voos para o modelo CL-600-2B19: "
```

```
## [1] 9588
```

```
## [1] "Percentual da quantidade de atrasos severos (> 90 minutos) para o modelo EMB-145XR: "
```

```
## [1] "16.04%"
```

```
## [1] "Percentual da quantidade de atrasos severos (> 90 minutos) para o modelo CL-600-2B19: "
```

```
## [1] "3.74%"
```

Analisando o impactos dos modelos EMB-145XR da Embraer e CL-600-2B19 da Bombardier para a companhia ExpressJet Airlines Inc, verifica-se que a contribuição percentual do modelo da Embraer nos atrasos severos (cerca de 16%) é bem maior que do modelo da Bombardier (cerca de 4%).

```
## [1] "Quadro com os atrasos totais para o modelo EMB-145XR"
```

```
## # A tibble: 1 x 4
##   model     total_atraso qtd_atraso atraso_medio
##   <chr>            <dbl>      <int>        <dbl>
## 1 EMB-145XR       587390       6403         91.7
```

```
## [1] "Quadro com os atrasos totais para o modelo CL-600-2B19"
```

```
## # A tibble: 1 x 4
##   model       total_atraso qtd_atraso atraso_medio
##   <chr>              <dbl>      <int>        <dbl>
## 1 CL-600-2B19       344076       3560         96.7
```

```
## [1] "Percentual da quantidade de atrasos totais (> 0 minuto) para o modelo EMB-145XR: "
```

```
## [1] "45.57%"
```

```
## [1] "Percentual da quantidade de atrasos totais (> 0 minuto) para o modelo CL-600-2B19: "
```

```
## [1] "37.13%"
```

Percentualmente, o impacto gerado pelo modelo CL-600-2B19 da Bombardier é menor que para o modelo da Embraer. O modelo da Bombardier atrasa cerca de 35% dos seus voos, enquanto o modelo da Embraer atrasa cerca 46% dos seus voos.

```
## [1] "Ganho percentual no total de atrasos em minutos com a mudança de modelo: "
```

```
## [1] "14.16%"
```

Dado que o percentual da quantidade de atrasos severos e também de atrasos gerais é menor para o modelo da Bombardier, será vantajoso, para a empresa que opera os aeroportos, que a ExpressJet Airlines Inc aposente o modelo da Embraer e passe a utilizar com mais frequência esse modelo da Bombardier. Esse ganho será em torno de 14% de economia no total de minutos de atraso.

```
## [1] "Quadro com os atrasos para a companhia JetBlue Airways"
```

```
## # A tibble: 7 x 5
##   model           total_atraso qtd_atraso atraso_medio desvioPad
##   <chr>                  <dbl>      <int>        <dbl>     <dbl>
## 1 A320-232              736401       3664         201.     111.
## 2 A321-231                2026         14         145.      31.8
## 3 ERJ 190-100 IGW       477575       2326         205.     116.
## 4 R66                     5854         32         183.      89.4
## 5 SR22                    4824         23         210.     129.
## 6 ZODIAC 601HDS           5482         30         183.      82.0
## 7 <NA>                   15529         82         189.     107.
```

```
## [1] "Quantidade de voos para o modelo A320-232: "
```

```
## [1] 45831
```

```
## [1] "Quantidade de voos para o modelo A321-231: "
```

```
## [1] 2878
```

```
## [1] "Percentual da quantidade de atrasos severos (> 90 minutos) para o modelo A320-232: "
```

```
## [1] "7.99%"
```

```
## [1] "Percentual da quantidade de atrasos severos (> 90 minutos) para o modelo A321-231: "
```

```
## [1] "0.03%"
```

Analisando o impactos dos modelos A320-232 e A321-231 da Airbus para a companhia JetBlue Airways, verifica-se que a contribuição percentual do modelo da A320-232 nos atrasos severos (cerca de 8%) é bem maior que do modelo da A321-231 (menor que 1%).

```
## [1] "Quadro com os atrasos totais para o modelo A320-232"
```

```
## # A tibble: 1 x 4
##   model    total_atraso qtd_atraso atraso_medio
##   <chr>           <dbl>      <int>        <dbl>
## 1 A320-232      1402334      19987         70.2
```

```
## [1] "Quadro com os atrasos totais para o modelo A321-231"
```

```
## # A tibble: 1 x 4
##   model    total_atraso qtd_atraso atraso_medio
##   <chr>           <dbl>      <int>        <dbl>
## 1 A321-231        61781       1078         57.3
```

```
## [1] "Percentual da quantidade de atrasos totais (> 0 minuto) para o modelo A320-232: "
```

```
## [1] "43.61%"
```

```
## [1] "Percentual da quantidade de atrasos totais (> 0 minuto) para o modelo A321-231: "
```

```
## [1] "37.46%"
```

Percentualmente, o impacto gerado pelo modelo A321-231 é menor que para o modelo A320-232. O modelo A321-231 atrasa cerca de 37% dos seus voos, enquanto o modelo A320-232 atrasa cerca 44% dos seus voos.

```
## [1] "Ganho percentual no total de atrasos em minutos com a mudança de modelo: "
```

```
## [1] "29.84%"
```

Dado que o percentual da quantidade de atrasos severos e também de atrasos gerais é menor para o modelo A321-231, será vantajoso, para a empresa que opera os aeroportos, que a JetBlue Airways aposente o modelo A320-232 e passe a utilizar com mais frequência esse modelo A321-231. Esse ganho será em torno de 30% de economia no total de minutos de atraso.

Segunda Preocupação:

> O aeroporto La Guardia é ideal para voos domésticos e de curta duração pois são mais cheios e aumentam consideravelmente a venda nas lojas do aeroporto. Pensando nisso, a diretoria quer proibir os voos com duração maior que três horas.

Agrupando o *Data Set* em dois grupos, um com voos menores e iguais a 180 minutos (3 horas) e outro com voos maiores que 180 minutos para o aeroporto de La Guardia.

```
flight_lga <- flight %>% filter(origin == 'LGA' | dest == 'LGA') %>%
  mutate(curta_dur = if_else(air_time <= 180, TRUE, FALSE),
         atraso_total = dep_delay + arr_delay) %>%
  select(year,month,day,carrier,tailnum,origin,dest,air_time,atraso_total,curta_dur)

paste("Percentual de NA's (air_time): ",
      round(flight_lga %>% filter(is.na(air_time) & origin == 'LGA') %>% nrow() /
              flight_lga %>% filter(origin == 'LGA') %>% nrow() * 100, 2), "%", sep = '')
```

```
## [1] "Percentual de NA's (air_time): 3.37%"
```

Verificando se o *Data Set* possui valores *NA* para o atributo *air_time*, identificou-se que esse percentual está em torno de 3%.

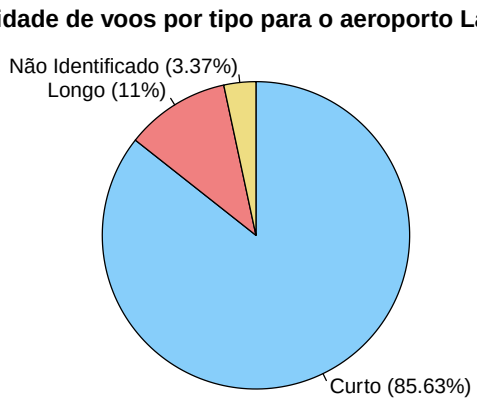
Avaliando os voos de acordo com o tipo (Curto, Longo e Não Identificado), nota-se que o aeroporto de La Guardia está bem concentrado em voos de curta duração (<= 180 minutos).

```
## [1] "Quadro com voos de acordo com o tipo para o aeroporto de La Guardia"
```

```
(flight_lga_tipo <- flight_lga %>% group_by(curta_dur) %>%
  summarise(qtd_voo = n(),
            media_hr_voo = mean(air_time)) %>%
  mutate(perc_qtd = qtd_voo / sum(qtd_voo),
         tipo_voo = case_when(curta_dur ~ "Curto",
                              curta_dur == FALSE ~ "Longo",
                              is.na(curta_dur) ~ "Não Identificado")) %>%
  arrange(desc(qtd_voo)) %>%
  select(tipo_voo,qtd_voo,perc_qtd,media_hr_voo))
```

```
## # A tibble: 3 x 4
##   tipo_voo         qtd_voo perc_qtd media_hr_voo
##   <chr>              <int>    <dbl>        <dbl>
## 1 Curto              89627   0.856          106.
## 2 Longo              11513   0.110          210.
## 3 Não Identificado    3523   0.0337          NA
```

```
pie(flight_lga_tipo$qtd_voo,
    labels = paste(flight_lga_tipo$tipo_voo, " (", round(flight_lga_tipo$perc_qtd * 100,2), "%)", sep =
    clockwise = TRUE, radius = 1,
    main = "Quantidade de voos por tipo para o aeroporto La Guardia",
    col = c('lightskyblue','lightcoral','lightgoldenrod'))
```



Verificando os voos de curta duração (<= 180 minutos) nos demais aeroportos de Nova Iorque, nota-se que os voos curtos são em grande maioria também.

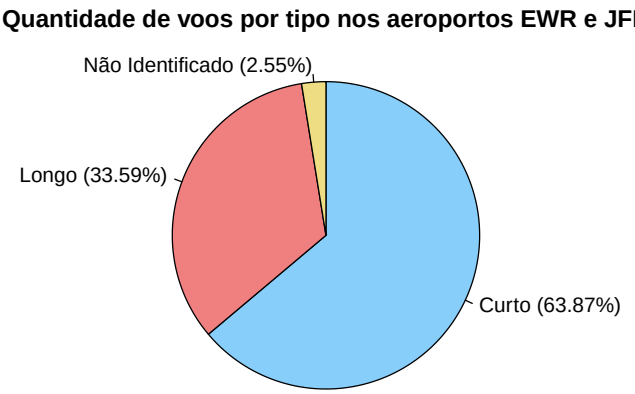
```
flight_ny <- flight %>% filter(origin %in% c('EWR','JFK') | dest %in% c('EWR','JFK')) %>%
  mutate(curta_dur = if_else(air_time <= 180, TRUE, FALSE),
         atraso_total = dep_delay + arr_delay) %>%
  select(year,month,day,carrier,tailnum,origin,dest,air_time,atraso_total,curta_dur)
```

```
## [1] "Quadro com voos de acordo com o tipo para os demais aeroportos de Nova Iorque (EWR e JFK)"
```

```
(flight_ny_tipo <- flight_ny %>% group_by(curta_dur) %>%
  summarise(qtd_voo = n(),
            media_hr_voo = mean(air_time)) %>%
  mutate(perc_qtd = qtd_voo / sum(qtd_voo),
         tipo_voo = case_when(curta_dur ~ "Curto",
                              curta_dur == FALSE ~ "Longo",
                              is.na(curta_dur) ~ "Não Identificado")) %>%
  arrange(desc(qtd_voo)) %>%
  select(tipo_voo,qtd_voo,perc_qtd,media_hr_voo))
```

```
## # A tibble: 3 x 4
##   tipo_voo         qtd_voo perc_qtd media_hr_voo
##   <chr>              <int>    <dbl>        <dbl>
## 1 Curto             148243   0.639         99.8
## 2 Longo              77963   0.336        290.
## 3 Não Identificado    5908   0.0255        NA
```

```
pie(flight_ny_tipo$qtd_voo,
    labels = paste(flight_ny_tipo$tipo_voo, " (", round(flight_ny_tipo$perc_qtd * 100,2), "%)", sep = "
    clockwise = TRUE, radius = 1,
    main = "Quantidade de voos por tipo nos aeroportos EWR e JFK",
    col = c('lightskyblue','lightcoral','lightgoldenrod'))
```



Observando o gráfica, percebe-se que existe a possibilidade de trocar os voos de longa duração no aerporto de La Guardia por voos de curta duração nos demais aeroportos de Nova Iorque.

Verificando a quantidade de voos que possuem tempo de voo maior que 180 minutos e menor que 190 minutos.

```
round(flight_lga %>% filter(!curta_dur & air_time > 180 & air_time <= 190) %>% nrow() /
  flight_lga %>% nrow() * 100, 2)
```

```
## [1] 2.06
```

Pecebe-se que somente cerca de 2% dos voos possuem duração entre 180 e 190 minutos. Portanto não há necessidade de flexibilizar o valor limite de 180 minutos para classificar um voo como curto.

Verificando os voos longos por companhia para o aerporto de La Guardia e os voos curtos para os demais aeroportos de Nova Iorque (Newark Liberty Intl e John F Kennedy Intl).

```
## [1] "Quadro com voos LONGOS por companhia para o aeroporto de La Guardia"
```

```
flight_lga %>% filter(!curta_dur) %>% group_by(carrier) %>%
  summarise(qtd_voo = n()) %>%
  mutate(perc_qtd = qtd_voo / sum(qtd_voo)) %>%
  inner_join(airline, by = c('carrier','carrier')) %>%
  arrange(name) %>%
  select(carrier,name,qtd_voo,perc_qtd)
```

```
## # A tibble: 7 x 4
##   carrier name                     qtd_voo perc_qtd
##   <chr>   <chr>                      <int>    <dbl>
## 1 AA      American Airlines Inc.      4175  0.363
## 2 DL      Delta Air Lines Inc.        1039  0.0902
## 3 9E      Endeavor Air Inc.             22  0.00191
## 4 MQ      Envoy Air                    231  0.0201
## 5 EV      ExpressJet Airlines Inc.      17  0.00148
## 6 F9      Frontier Airlines Inc.       681  0.0592
## 7 B6      JetBlue Airways               66  0.00573
```

```
## [1] "Quadro com voos CURTOS por companhia para os demais aeroportos de Nova Iorque (EWR e JFK)"
```

```
flight_ny %>% filter(curta_dur) %>% group_by(carrier) %>%
  summarise(qtd_voo = n()) %>%
  mutate(perc_qtd = qtd_voo / sum(qtd_voo)) %>%
  inner_join(airline, by = c('carrier','carrier')) %>%
  arrange(name) %>%
  select(carrier,name,qtd_voo,perc_qtd)
```

```
## # A tibble: 6 x 4
##   carrier name                     qtd_voo perc_qtd
##   <chr>   <chr>                      <int>    <dbl>
## 1 AA      American Airlines Inc.      6645   0.0448
## 2 DL      Delta Air Lines Inc.       11727   0.0791
## 3 9E      Endeavor Air Inc.          14496   0.0978
## 4 MQ      Envoy Air                   8933   0.0603
## 5 EV      ExpressJet Airlines Inc.   42061   0.284
## 6 B6      JetBlue Airways            33938   0.229
```

Percebe-se que temos 7 companhias que serão afetadas com a proibição de voos longos no aerporto de La Guardia, sendo que 6 delas, American Airlines Inc, Delta Air Lines Inc, Endeavor Air Inc, Envoy Air, ExpressJet Airlines Inc e JetBlue Airways, possuem voos curtos nos demais aeroportos de Nova Iorque. Somente uma, Frontier Airlines Inc, não possui essa característica.

Conclusão:

> Cerca de 85% dos voos que partem do aeroporto de La Guardia são voos de curta duração, assim o impacto inicial em proibir voos com tempo de duração maiores que 180 minutos não irá causar transtornos significativos para as companhias.
> Serão 7 companhias que possuem voos longos no aerporto de La Guardia que serão impactadas. Sendo que 6 delas, American Airlines Inc, Delta Air Lines Inc, Endeavor Air Inc, Envoy Air, ExpressJet Airlines Inc e JetBlue Airways, possuem voos curtos nos demais aeroportos de Nova Iorque (Newark Liberty Intl e John F Kennedy Intl) e poderão trocar os voos longo por curto, passando a operar somente com voos curtos no La Guardia.
> A Frontier Airlines Inc é a única companhia que não terá a possiblidade de efetuar essa troca, pois não possui voos de curta duração nos demais aeroportos de Nova Iorque. Portanto ela terá que transferir suas operações com voos longos para os demais aeroportos de Nova Iorque.

Terceira Preocupação:

> A área de manutenção e verificação das aeronaves deseja analisar a participação de cada fabricante de aeronaves nos aeroportos da cidade de Nova Iorque. Assim, a área precisa da relação entre as informações de cada voo e o fabricante do avião.

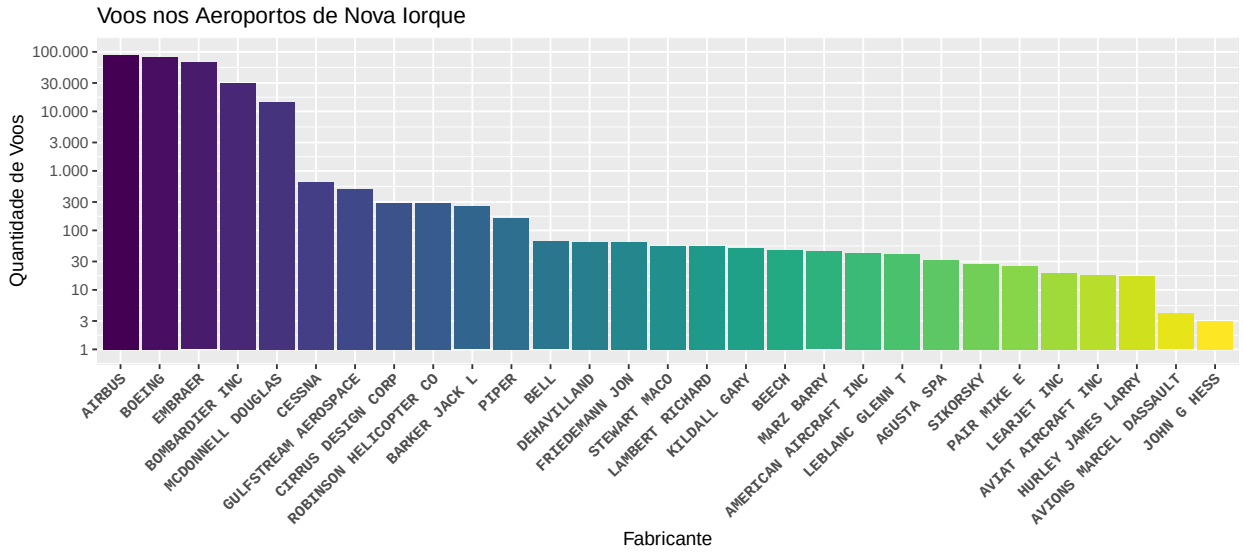
Criando um *Data Set* com os dados de voos que saem e chegam nos aeroportos de Nova Iorque por fabricante.

```
flight_plane <- flight %>% filter(origin %in% c('LGA','JFK','EWR') | dest %in% c('LGA','JFK','EWR')) %>%
  select(year,month,day,carrier,flight,tailnum,origin,dest,distance)

flight_plane <- flight_plane %>% inner_join(plane, by = c('tailnum','tailnum')) %>%
  mutate(year_flight = year.x, year_model = year.y) %>%
```

```
  select(year_flight,month,day,carrier,flight,tailnum,manufacturer,model,year_model,origin,dest,distanc

flight_plane_agg <- flight_plane %>% group_by(manufacturer) %>%
  summarise(distance = sum(distance),
            qtd_voo = n()) %>%
  arrange(desc(qtd_voo))
```

Ordenando as informações pela quantidade de voos.

```
flight_plane_agg_ord <- flight_plane_agg
indice <- order(flight_plane_agg$qtd_voo, decreasing = TRUE)
nivel <- flight_plane_agg$manufacturer[indice]
flight_plane_agg_ord$manufacturer <- factor(flight_plane_agg_ord$manufacturer, levels = nivel, ordered =

flight_plane_agg_ord %>%
  ggplot(aes(x = manufacturer, y = qtd_voo)) +
  geom_bar(aes(fill = flight_plane_agg_ord$manufacturer), stat = "identity", show.legend = FALSE) +
  scale_y_log10(limits = c(1,100000),
                breaks = c(1,3,10,30,100,300,1e3,3e3,1e4,3e4,1e5),
                labels = c('1','3','10','30','100','300','1.000','3.000','10.000','30.000','100.000')) +
  labs(x = "Fabricante", y = "Quantidade de Voos", title = "Voos nos Aeroportos de Nova Iorque") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, family = 'mono', face = 'bold'))
```

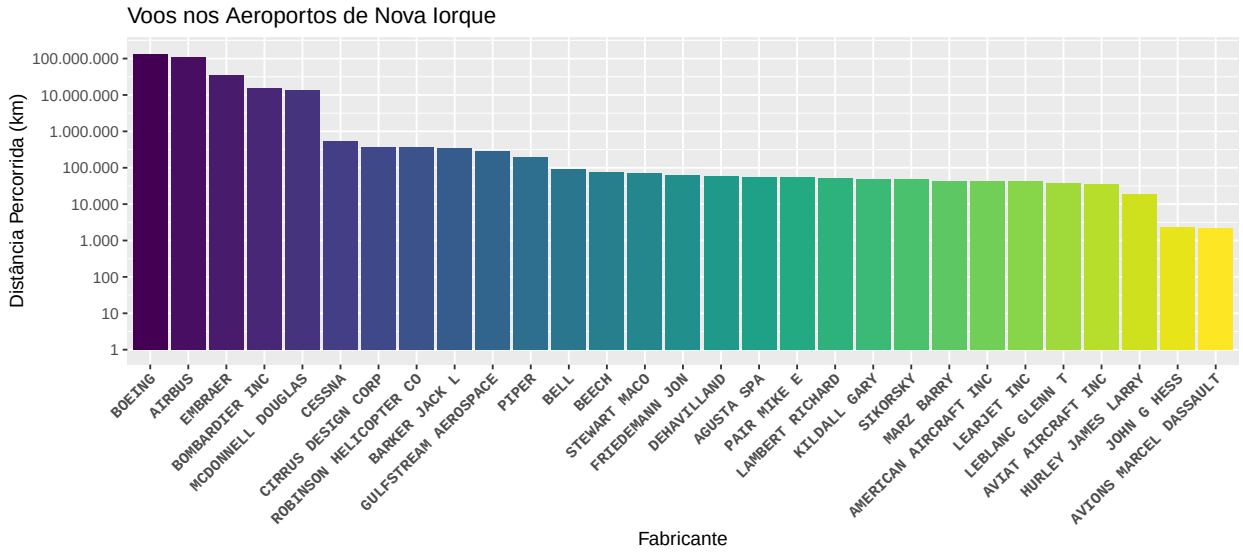


Percebe-se uma concentração da quantidade de voos em 5 fabricantes: Airbus, Boeing, Bambardier Inc, Embraer e McDonnell Douglas.

Mudando o atributo de ordenação para distância percorrida em Kilômetro.

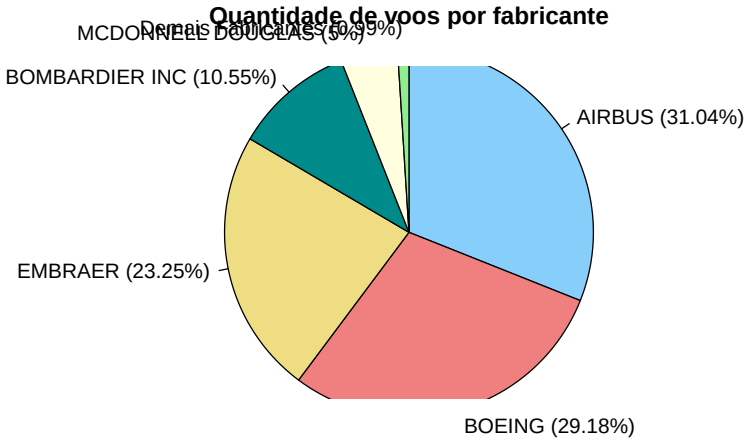
```
flight_plane_agg_ord <- flight_plane_agg
indice <- order(flight_plane_agg$distance, decreasing = TRUE)
nivel <- flight_plane_agg$manufacturer[indice]
flight_plane_agg_ord$manufacturer <- factor(flight_plane_agg_ord$manufacturer, levels = nivel, ordered =

flight_plane_agg_ord %>%
  ggplot(aes(x = manufacturer, y = distance)) +
```

```
  geom_col(aes(fill = flight_plane_agg_ord$manufacturer), show.legend = FALSE) +
  scale_y_log10(limits = c(1,1.5e8),
                breaks = c(1,10,100,1000,10000,100000,1000000,10000000,100000000),
                labels = c('1','10','100','1.000','10.000','100.000','1.000.000','10.000.000','100.000.
  labs(x = "Fabricante", y = "Distância Percorrida (km)", title = "Voos nos Aeroportos de Nova Iorque")
  theme(axis.text.x = element_text(angle = 45, hjust = 1, family = 'mono', face = 'bold'))
```



Nota-se que a concentração nos 5 fabricantes permanece.

```
flight_plane_agg_aux <- flight_plane_agg %>%
  filter(manufacturer %in% c('BOEING','AIRBUS','EMBRAER','BOMBARDIER INC','MCDONNELL DOUGLAS')) %>%
  union(
    flight_plane_agg %>%
      filter(!manufacturer %in% c('BOEING','AIRBUS','EMBRAER','BOMBARDIER INC','MCDONNELL DOUGLAS')) %>%
      summarise(manufacturer = 'Demais Fabricantes',
                distance = sum(distance),
                qtd_voo = sum(qtd_voo)))
flight_plane_agg_aux <- flight_plane_agg_aux %>% mutate(perc_voo = qtd_voo / sum(qtd_voo))

pie(flight_plane_agg_aux$qtd_voo,
    labels = paste(flight_plane_agg_aux$manufacturer, " (",
                   round(flight_plane_agg_aux$perc_voo * 100, 2),
                   "%)", sep = ''),
    clockwise = TRUE,
    radius = 1.2,
    main = "Quantidade de voos por fabricante",
    col = c('lightskyblue','lightcoral','lightgoldenrod','cyan4','lightyellow','lightgreen','lightpink'
```



Conclusão:

Existe uma concentração grande de voos nos fabricantes Airbus, Boeing, Bambardier Inc, Embraer e McDonnell Douglas para os aeroportos de Nova Iorque (La Guardia, Newark Liberty e John F Kennedy). Essa concentração se dá, quando a análise é feita utilizando a quantidade de voos e pemanece a mesma, quando é utilizado a distância percorrida em Kilômetro.
Esses 5 fabricantes respondem por quase a totalidade dos voos nesses aeroportos, sendo que a Airbus e a Boeing respondem por mais da metada desses voos.
Isso é esperado, dado que são as duas maiores fabricantes de aviões no mundo.

Quarta Preocupação:

A área financeira sabe que operar aviões maiores e com mais motores é mais caro e que por isso, esses aviões deveriam ter mais assentos, contudo ela acha que as empresas aéreas não entenderam muito bem isso e que essa relação (total de assentos / total de motores) não está clara.

Agrupando os dados em um *Data Set* para análise.

```
flight_plane_airline <- flight %>% select(tailnum,carrier,flight) %>% distinct() %>%
  left_join(plane, by = c('tailnum','tailnum')) %>%
  select(tailnum,manufacturer,model,engines,seats,carrier) %>%
  left_join(airline, by = c('carrier','carrier')) %>%
  select(tailnum,manufacturer,model,engines,seats,carrier,name) %>%
  arrange(manufacturer,model)
```

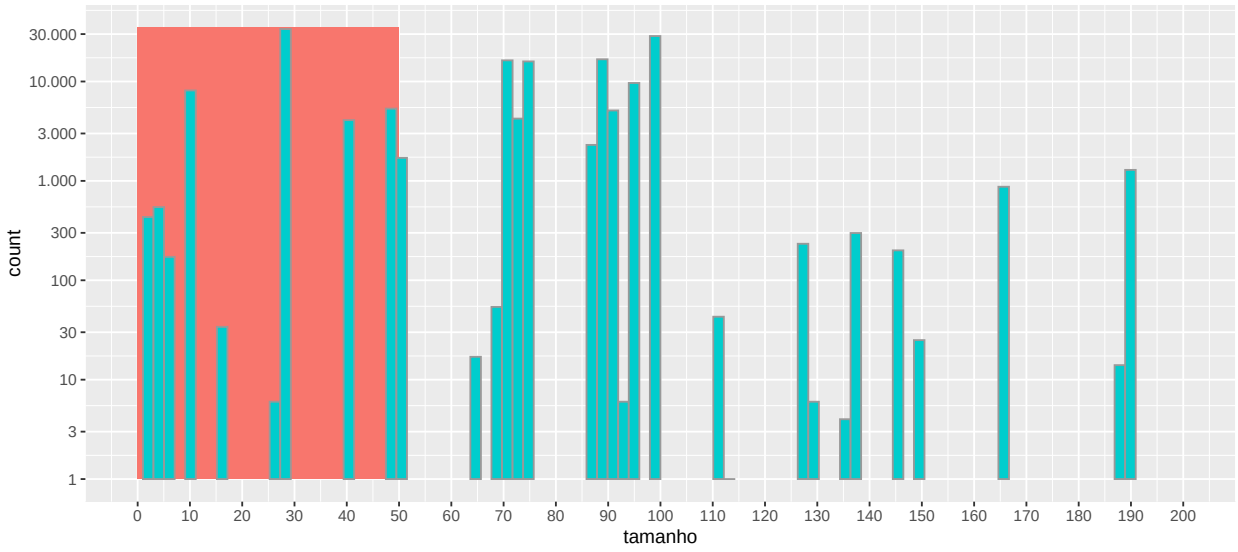
Verificando a ausência de informação relevante para a análise e calculando a relação $\frac{TotaldeAssentos}{TotaldeMotores}$.

```
paste("Percentual de NA's nas colunas seats e engines: ",
      round(sum(if_else(is.na(flight_plane_airline$seats) | is.na(flight_plane_airline$engines),TRUE,FA
              flight_plane_airline %>% nrow() * 100, 2), '%', sep = '')
```

```
## [1] "Percentual de NA's nas colunas seats e engines: 13.17%"
```

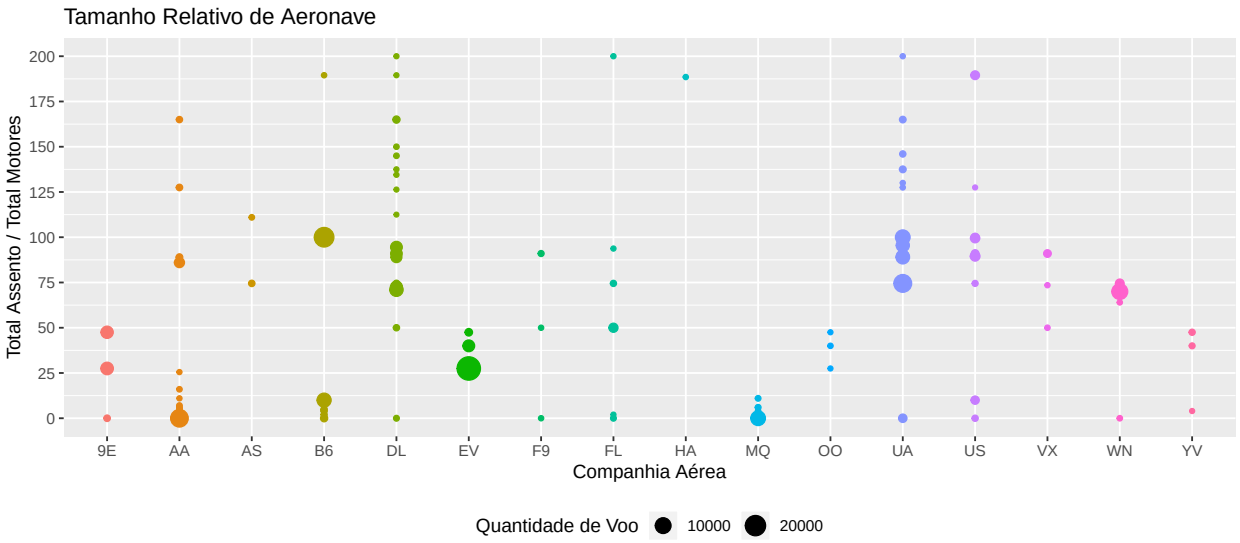
```
# Calculando o Tamanho Relativo das Aeronaves: Total Assentos / Total Motores
flight_plane_airline <- flight_plane_airline %>%
  mutate(tamanho = if_else(is.na(seats),0,as.double(seats)) / if_else(is.na(engines),1,as.double(engines
```

```
flight_plane_airline %>%
  ggplot(aes(tamanho)) +
  geom_rect(aes(xmin = 0 , xmax = 50, ymin = 1, ymax = 35000, fill = 'red'), show.legend = FALSE) +
  geom_histogram(bins = 100, fill = 'cyan3', color = 'gray60') +
  scale_x_continuous(limits = c(0,200),
                     breaks = c(0,10,20,30,40,50,60,70,80,90,100,110,120,130,140,150,160,170,180,190,200
                     labels = c('0','10','20','30','40','50','60','70','80','90','100','110','120','130
                                '140','150','160','170','180','190','200')) +
  scale_y_log10(limits = c(1,35000),
                breaks = c(1,3,10,30,100,300,1000,3000,10000,30000),
                labels = c('1','3','10','30','100','300','1.000','3.000','10.000','30.000'))
```



Considerando uma relação Total de Assentos / Total de Motores abaixo de 50 como desfavorável, pois teríamos menos que cinquenta passageiros por motor, percebe-se que existe uma quantidade razoável de voos abaixo desse nível. Isso mostra que temos várias empresas que ainda utilizam aeronaves pouco rentáveis.

```
flight_plane_airline %>% group_by(carrier, tamanho) %>% summarise(qtd = n()) %>%
  left_join(airline, by = c('carrier', 'carrier')) %>%
  select(carrier,name,tamanho,qtd) %>%
  ggplot(aes(x = carrier, y = tamanho)) +
  geom_point(aes(size = qtd, color = carrier)) +
  scale_y_continuous(limits = c(0,200),
                     breaks = c(0,25,50,75,100,125,150,175,200)) +
  guides(color = 'none') +
  labs(x = "Companhia Aérea", y = "Total Assento / Total Motores",
       size = "Quantidade de Voo", title = "Tamanho Relativo de Aeronave") +
  theme(legend.position = "bottom")
```



Investigando mais detalhadamente por companhia aérea, percebe-se que as companhias Endeavor Air Inc (9E), American Airlines (AA), JetBlue Airways (B6), ExpressJet Airlines Inc (EV) e Envoy Air (MQ) possuem uma grande quantidade de voos com uma baixa relação Total de Assentos / Total de Motores.

```
paste("Percentual de NA's nos dados relavantes para a empresa American Airlines (AA): ",
      round(flight_plane_airline %>% filter(carrier == 'AA' & tamanho == 0) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'AA') %>% nrow() * 100, 2),
      "%", sep = '')
```

```
## [1] "Percentual de NA's nos dados relavantes para a empresa American Airlines (AA): 79.19%"
```

```
paste("Percentual de NA's nos dados relavantes para a empresa Envoy Air (MQ): ",
      round(flight_plane_airline %>% filter(carrier == 'MQ' & tamanho == 0) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'AA') %>% nrow() * 100, 2),
      "%", sep = '')
```

```
## [1] "Percentual de NA's nos dados relavantes para a empresa Envoy Air (MQ): 46.65%"
```

Levantando os percentuais de NA's nos dados relevantes para as companhias American Airlines (AA) e Envoy Air (MQ), observa-se que a posição dessas duas companhias no gráfico *Tamanho Relativo de Aeornave* não é porque elas estão operando voos com aeronaves de relação desfavorável, ou seja, com relação Total de Assentos / Total de Motores menor que 50, e sim que não há dados disponíveis para o cálculo adequado dessa relação. Portanto, não há como obtermos qualquer conclusão a respeito dessas duas companhias.

```
paste("Percentual de NA's nos dados relavantes para a empresa JetBlue Airways (B6): ",
      round(flight_plane_airline %>% filter(carrier == 'B6' & tamanho == 0) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'B6') %>% nrow() * 100, 2),
      "%", sep = "")
```

```
## [1] "Percentual de NA's nos dados relavantes para a empresa JetBlue Airways (B6): 1.67%"
```

```
paste("Percentual de Voos com Relação Desfavorável para a empresa JetBlue Airways (B6): ",
      round(flight_plane_airline %>% filter(carrier == 'B6' & tamanho <= 50) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'B6') %>% nrow() * 100, 2),
      "%", sep = "")
```

```
## [1] "Percentual de Voos com Relação Desfavorável para a empresa JetBlue Airways (B6): 30.52%"
```

```
paste("Percentual de NA's nos dados relavantes para a empresa Endeavor Air Inc (9E): ",
      round(flight_plane_airline %>% filter(carrier == '9E' & tamanho == 0) %>% nrow() /
              flight_plane_airline %>% filter(carrier == '9E') %>% nrow() * 100, 2),
      "%", sep = "")
```

```
## [1] "Percentual de NA's nos dados relavantes para a empresa Endeavor Air Inc (9E): 2.29%"
```

```
paste("Percentual de Voos com Relação Desfavorável para a empresa Endeavor Air Inc (9E): ",
      round(flight_plane_airline %>% filter(carrier == '9E' & tamanho <= 50) %>% nrow() /
              flight_plane_airline %>% filter(carrier == '9E') %>% nrow() * 100, 2),
      "%", spe = "")
```

```
## [1] "Percentual de Voos com Relação Desfavorável para a empresa Endeavor Air Inc (9E):  100 % "
```

```
paste("Percentual de NA's nos dados relavantes para a empresa ExpressJet Airlines Inc (EV): ",
      round(flight_plane_airline %>% filter(carrier == 'EV' & tamanho == 0) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'EV') %>% nrow() * 100, 2),
      "%", sep = "")
```

```
## [1] "Percentual de NA's nos dados relavantes para a empresa ExpressJet Airlines Inc (EV): 0%"
```

```
paste("Percentual de Voos com Relação Desfavorável para a empresa ExpressJet Airlines (EV): ",
      round(flight_plane_airline %>% filter(carrier == 'EV' & tamanho <= 50) %>% nrow() /
              flight_plane_airline %>% filter(carrier == 'EV') %>% nrow() * 100, 2),
      "%", sep = "")
```

```
## [1] "Percentual de Voos com Relação Desfavorável para a empresa ExpressJet Airlines (EV): 100%"
```

Investigando mais detalhadamente os dados para as companhias JetBlue Airways (B6), Endeavor Air Inc (9E) e ExpressJet Airlines Inc (EV), percebe-se que os NA's não explicam suas posições no gráfico *Tamanho Relativo de Aeornave*. Observando o *Percentual de Voos com Relação Desfavorável* para essas companhias, notamos que a relação desfavorável é causada pela operação de voos com aeronaves menores, ou seja, com relação Total de Assentos / Total de Motores menor que 50.

Conclusão:

As companhias Endeavor Air Inc (9E) e ExpressJet Airlines Inc (EV) estão operando somente com aeronaves que possuem uma relação Total de Assentos / Total de Motores desfavorável. Isso é evidenciado pelo *Percentual de Voos com Relação Desfavorável*, que para essas empresas geraram o valor de 100%.
A companhia JetBlue Airways (B6) opera com 30% de seus voos com aeronaves que possuem uma relação Total de Assentos / Total de Motores desfavorável, isso também é evidenciado pelo *Percentual de Voos com Relação Desfavorável* para essa companhia.

Quinta Preocupação:

Uma grande preocupação na área da aviação é a visibilidade do piloto. Sabe-se que os estudos sobre as condições climáticas influenciam diretamente nessa variável e "a diretoria" gostaria de entender melhor essa relação.