Data Visualization Homework

Thamonwan Nuchtisan

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(glue)
library(nycflights13)
library(patchwork)
library(lubridate)
```

```
## Loading required package: timechange
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggthemes)
library(dplyr)
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggtext)
library(RColorBrewer)
```

Chart 1 : Bubble chart

Relationship between airline and fuel expenses

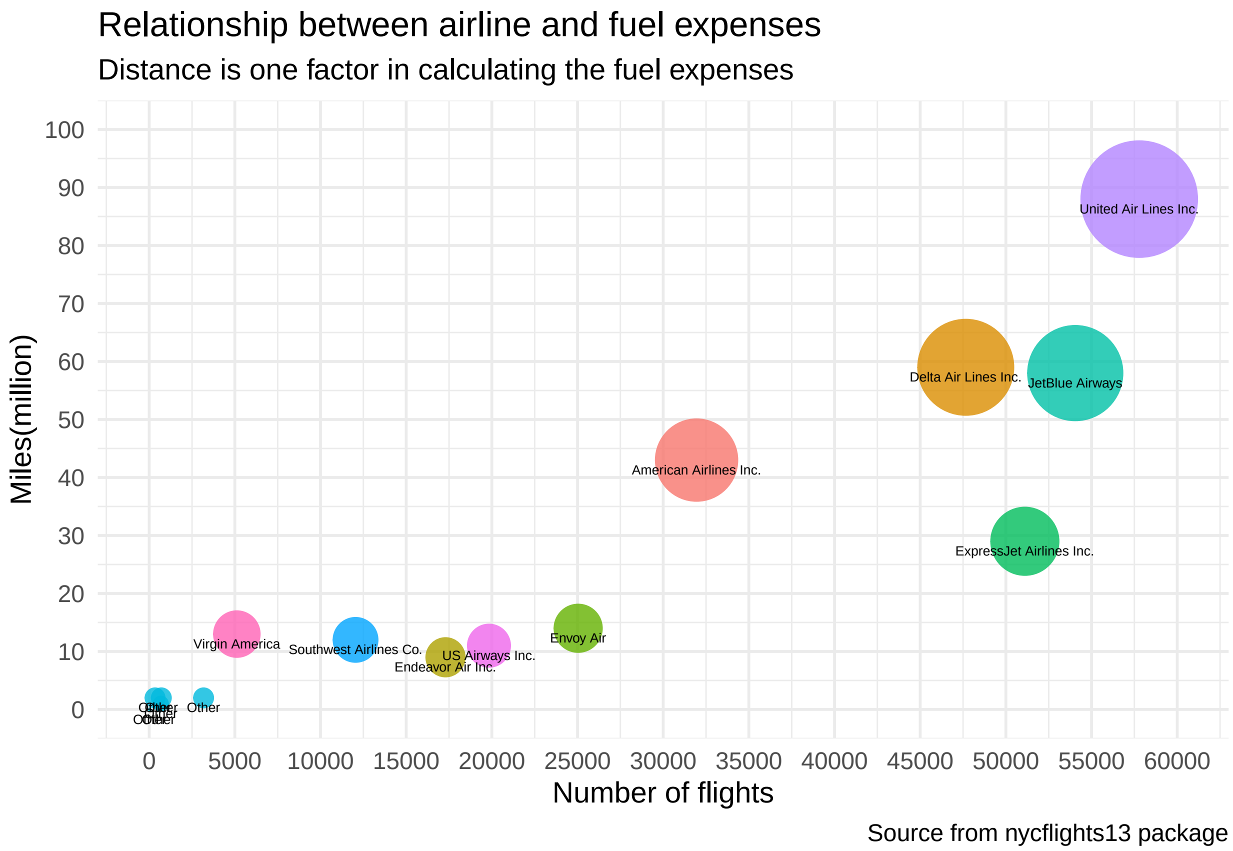
```
##Prepare Data
# Find cancelled flights
not_cancelled <- filter(flights,!is.na(arr_time),!is.na(air_time))

# Find amount of miles from distance column
df1 <- not_cancelled %>%
  group_by(carrier) %>%
  summarise(miles = sum(distance),
            count = n()) %>%
  mutate(flight_rank=round(count/sum(count)*100,digits = 2)) %>%
  mutate(mile_rank=round(miles/sum(miles)*100,digit = 2)) %>%
  mutate(miles = round(miles/1000000),digit = 2) %>%
  arrange(desc(miles))

df2 <- left_join(df1,airlines, by = "carrier") %>%
  mutate(name = if_else(mile_rank < 1, "Other", name))
```

```
# Find top 5 of airline it probably high fuel costs.
# If distance is one factor in calculating the fuel expenses
knitr::kable(
  left_join(df1,airlines, by = "carrier") %>%
  select(carrier,name,miles) %>%
  rename("miles(million)" = "miles") %>%
  head(5),
  caption = "The top 5 of airlines may have high fuel costs")
```

Table 1: The top 5 of airlines may have high fuel costs

carrier	name	miles(million)
UA	United Air Lines Inc.	88
DL	Delta Air Lines Inc.	59
B6	JetBlue Airways	58
AA	American Airlines Inc.	43
EV	ExpressJet Airlines Inc.	29

```r
## Plot graph by ggplot2
# Bubble chart
ggplot(data = df2, mapping = aes(x=count,
                                  y=miles,
                                  size=miles,
                                  color=name)) +
  geom_point(alpha=0.8) +
  scale_size_area(max_size = 20) +
  geom_text(aes(label=name),size=1.8,color = "black",hjust = 0.5, vjust = 1.5) +
  scale_y_continuous(limits = c(0, 100),
                     breaks = seq(0, 100, by=10)) +
  scale_x_continuous(limits = c(0,60000),
                     breaks = seq(0,60000, by=5000)) +
  theme_minimal() +
  guides(size="none",
         color="none") +
  labs(
    x = "Number of flights",
    y = "Miles(million)",
    title = "Relationship between airline and fuel expenses ",
    subtitle = "Distance is one factor in calculating the fuel expenses",
    caption = "Source from nycflights13 package")
```



Relationship between airline and fuel expenses

Distance is one factor in calculating the fuel expenses

Source from nycflights13 package

Chart 2 : Bar chart

Relationship between Not cancelled flights and Cancelled flights for each carrier

```
## Prepare Data
# mutate column NorC separate data Not cancelled flight and Cancelled flight
f1 <- flights %>%
  mutate(NorC = factor(if_else(!is.na(arr_delay) & !is.na(air_time),"notcancelled","cancelled"),
                       levels = c("cancelled","notcancelled"),
                       labels = c("cancelled","notcancelled"),))
p1 <- f1 %>%
  filter(NorC == "cancelled") %>%
  group_by(carrier,NorC) %>%
  count(NorC, name = "count_cancelled") %>%
  select(carrier,NorC,count_cancelled) %>%
  arrange(NorC,desc(count_cancelled)) %>%
  ungroup()
# Find Top 5 Domestic Carriers (by flights count)
p2 <- f1 %>%
  filter(NorC == "notcancelled") %>%
  group_by(carrier,NorC) %>%
  count(NorC, name = "count_notcancelled") %>%
  select(carrier,NorC,count_notcancelled) %>%
  arrange(NorC,desc(count_notcancelled)) %>%
  head(5) %>%
  ungroup()

# Find the percentage of canceled flights for each of the top 5 domestic carriers.
j1 <- left_join(p2,p1, by = "carrier")
knitr::kable(
  left_join(j1,airlines, by = "carrier") %>%
  mutate(percent_cancelled = round((count_cancelled/count_notcancelled)*100,digit = 2)) %>%
  mutate(percent_cancelled = paste0(percent_cancelled," %")) %>%
  select(carrier,name,count_notcancelled,count_cancelled,percent_cancelled),
  caption = "The percentage of canceled flights for each of the top 5 domestic carriers")
```

Table 2: The percentage of canceled flights for each of the top 5 domestic carriers

carrier	name	count_notcancelled	count_cancelled	percent_cancelled
UA	United Air Lines Inc.	57782	883	1.53 %
B6	JetBlue Airways	54049	586	1.08 %
EV	ExpressJet Airlines Inc.	51108	3065	6 %
DL	Delta Air Lines Inc.	47658	452	0.95 %
AA	American Airlines Inc.	31947	782	2.45 %

```
# Find amount of  Not cancelled flights and Cancelled for  each carrier
p3 <- f1 %>%
  group_by(carrier,NorC) %>%
  count(NorC, name = "count") %>%
  select(carrier,NorC,count) %>%
  arrange(NorC,desc(count)) %>%
  ungroup()
```

```
## Plot graph by ggplot2
# Bar chart
ggplot(data = p3, mapping = aes(x = carrier,
                                y = count,
                                fill = NorC)) +
  geom_bar(aes(y = count),stat = "identity") + ## stat "identity" own value
  scale_fill_manual(values = c("cancelled" = "#eb4d4d", "notcancelled" = "#310d63" )) +
  labs(
    y = "Number of flights",
    title = "Relationship between Not cancelled flights and Cancelled flights for each carrier",
    caption = "Source from nycflights13 package"
  )
```

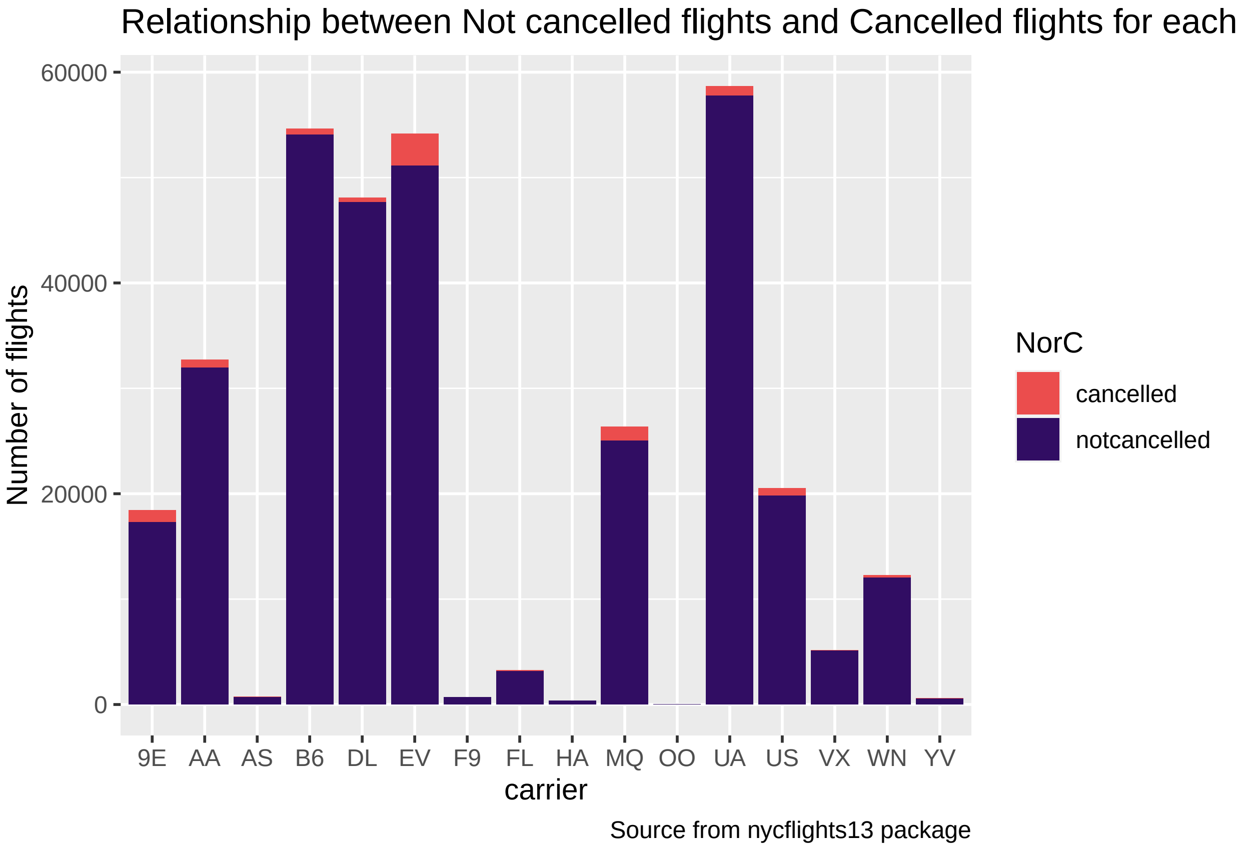


Chart 3 : Bar chart

Relationship between top 5 manufacturers and age of aircraft in service

```
##Prepare Data
# Find cancelled flights and clear missing data of tail number column
d1 <- flights %>%
  filter(is.na(arr_delay),is.na(air_time),!is.na(tailnum)) %>%
  select(carrier,flight,tailnum,origin,dest)

# Find top 5 aircraft manufacturers of cancelled flights data
d2 <- left_join(d1,planes, by = "tailnum")
knitr::kable(
  d2 %>%
  filter(!is.na(manufacturer)) %>%
  group_by(manufacturer) %>%
  count(manufacturer, sort = TRUE) %>%
  head(5),
  caption = "Top 5 aircraft manufacturers of cancelled flights data")
```

Table 3: Top 5 aircraft manufacturers of cancelled flights data

manufacturer	n
EMBRAER	2538
BOMBARDIER INC	842
BOEING	629
AIRBUS	412
AIRBUS INDUSTRIE	249

```
# Find the age of the aircraft that is still in service
# Determine average of aircaft. Assume average of aircaft is around 20 year
d3 <- d2 %>% filter(!is.na(year)) %>%
  mutate(duration = 2013 - year) %>%
  mutate(age_plane = if_else(duration < 20 , "lower 20 yrs", "since 20 yrs up")) %>%
  mutate(age_plane = factor(age_plane,
                            levels = c("lower 20 yrs", "since 20 yrs up"),
                            labels = c("lower 20 yrs", "since 20 yrs up"),
                            ordered = TRUE )) %>%
  filter(manufacturer %in% c("BOEING",
                             "EMBRAER",
                             "BOMBARDIER INC",
                             "AIRBUS",
                             "AIRBUS INDUSTRIE" )) # top 5 aircraft manufacturers

d4 <- d3 %>%
  group_by(manufacturer,age_plane) %>%
  count(manufacturer, sort = T)
```

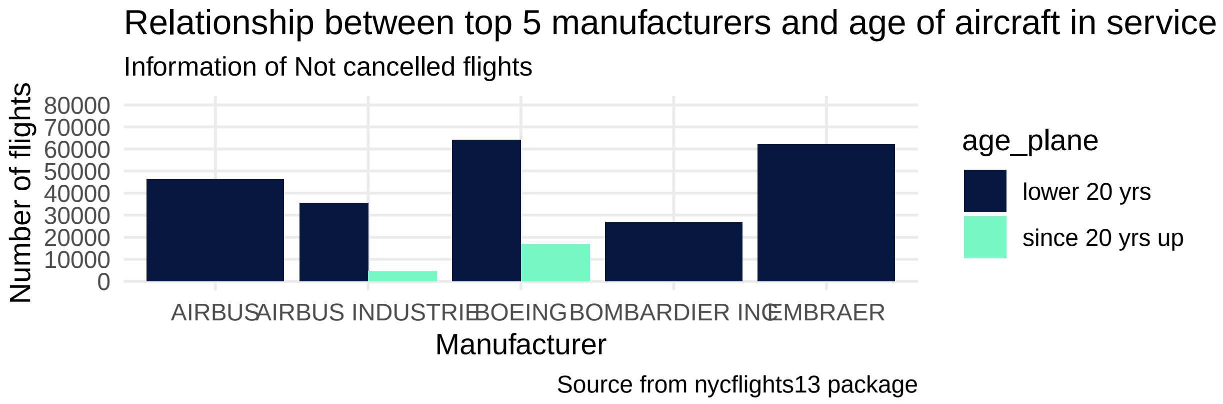
```r
# Find Not cancelled flights and clear missing data of tail number column
e1 <- flights %>% filter(!is.na(arr_delay),!is.na(air_time),!is.na(tailnum)) %>%
  select(carrier,flight,tailnum,origin,dest)

# Find top 5 aircraft manufacturers of Not cancelled flights data
e2 <- left_join(e1,planes, by = "tailnum")
knitr::kable(
  e2 %>%
  filter(!is.na(manufacturer)) %>%
  group_by(manufacturer) %>%
  count(manufacturer, sort = TRUE) %>%
  head(5),
  caption = "Top 5 aircraft manufacturers of Not cancelled flights data")
```

Table 4: Top 5 aircraft manufacturers of Not cancelled flights data

manufacturer	n
BOEING	82283
EMBRAER	63530
AIRBUS	46890
AIRBUS INDUSTRIE	40642
BOMBARDIER INC	27430

```r
# Find the age of the aircraft that is still in service
# Determine average of aircaft. Assume average of aircaft is around 20 year
e3 <- e2 %>% filter(!is.na(year)) %>%
  mutate(duration = 2013 - year) %>%
  mutate(age_plane = if_else(duration < 20 , "lower 20 yrs", "since 20 yrs up")) %>%
  mutate(age_plane = factor(age_plane,
                            levels = c("lower 20 yrs", "since 20 yrs up"),
                            labels = c("lower 20 yrs", "since 20 yrs up"),
                            ordered = TRUE )) %>%
  filter(manufacturer %in% c("BOEING",
                             "EMBRAER",
                             "BOMBARDIER INC",
                             "AIRBUS",
                             "AIRBUS INDUSTRIE" )) # top 5 aircraft manufacturers

e4 <- e3 %>%
  group_by(manufacturer,age_plane) %>%
  count(manufacturer, sort = T)
```

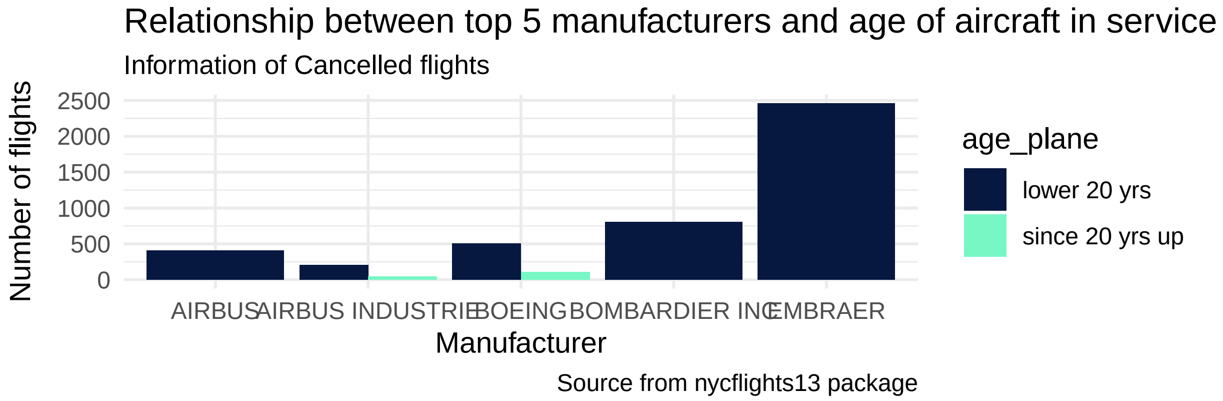
```r
## Plot graph by ggplot2
# Not cancelled flights
g1 <- ggplot(data = e4, mapping = aes(x = manufacturer,
                                      y = n,
                                      fill = age_plane)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_fill_manual(values = c("lower 20 yrs" = "#061740", "since 20 yrs up" = "#77f7c4" )) +
  scale_y_continuous(limits = c(0, 80000),
                     breaks = seq(0, 80000, by=10000),
                     minor_breaks = NULL ) +
  theme_minimal() +
  theme(plot.subtitle = element_text(size = 10)) +
  labs(
    x = "Manufacturer",
    y = "Number of flights",
    title = "Relationship between top 5 manufacturers and age of aircraft in service",
    subtitle = "Information of Not cancelled flights",
    caption = "Source from nycflights13 package"
  )

# Cancelled flights
g2 <- ggplot(data = d4, mapping = aes(x = manufacturer,
                                      y = n,
                                      fill = age_plane)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_fill_manual(values = c("lower 20 yrs" = "#061740", "since 20 yrs up" = "#77f7c4" )) +
  theme_minimal() +
  theme(plot.subtitle = element_text(size = 10)) +
  labs(
    x = "Manufacturer",
    y = "Number of flights",
    title = "Relationship between top 5 manufacturers and age of aircraft in service",
    subtitle = "Information of Cancelled flights",
    caption = "Source from nycflights13 package"
  )
g1 / g2
```



Relationship between top 5 manufacturers and age of aircraft in service

Information of Not cancelled flights

Source from nycflights13 package



Relationship between top 5 manufacturers and age of aircraft in service

Information of Cancelled flights

Source from nycflights13 package

Chart 4 : Line chart

Flights per day of 3 Departure Airport

```
##Prepare Data
# No cancelled flights
df_nc <- filter(flights,!is.na(arr_time),!is.na(air_time))

# mutate column date from column time_hour
df_nc <- df_nc %>%
  mutate(date = date(time_hour))

# count flights per day
df1 <- df_nc %>%
  group_by(date, origin) %>%
  count(origin) %>%
  arrange(origin)
# Find average flights per day of each 3 airport
df2 <- df1 %>%
  group_by(origin) %>%
  summarise(avg_per_day = round(mean(n),digits = 0)) %>%
  ungroup()

knitr::kable(
  left_join(df2,airports, by = c("origin" = "faa")) %>%
  select(origin,name,avg_per_day),
  caption = "The average flights per day of each 3 airport")
```

Table 5: The average flights per day of each 3 airport

origin	name	avg_per_day
EWR	Newark Liberty Intl	321
JFK	John F Kennedy Intl	299
LGA	La Guardia	277

```
## Plot graph by ggplot2
# For create average line
al <- aggregate(n~origin, df1, FUN = mean, na.rm = T)
# Line chart
ggplot(data = df1, mapping = aes(x = date,
                                 y = n,
                                 group = factor(origin),
                                 color = origin)) +
  geom_line(linewidth = 0.4, alpha = 0.8) +
  geom_point(size = 0.2) +
  facet_grid(origin~.) +
  geom_vline(aes(xintercept = as.Date("2013-03-31")),
             color = "black", linetype = 2, show.legend = F) +
  geom_vline(aes(xintercept = as.Date("2013-06-30")),
             color = "black", linetype = 2, show.legend = F) +
  geom_vline(aes(xintercept = as.Date("2013-09-30")),
             color = "black", linetype = 2, show.legend = F) +
  geom_vline(aes(xintercept = as.Date("2013-12-31")),
             color = "black", linetype = 2, show.legend = F) +
  geom_hline(aes(yintercept = as.numeric(n)), data = al ,
             color = "black", linetype = "dashed") +
  scale_x_date(limits = c(as.Date("2013-01-01"),as.Date("2013-12-31")),
               date_labels = '%b %Y',
               breaks = as.Date(c("2013-03-31", "2013-06-30", "2013-09-30", "2013-12-31"))) +
  scale_color_manual(values = c("EWR" = "#CB4335",
                                "JFK" = "#dbc323",
                                "LGA" = "#5DADE2")) +
  theme_minimal () +
  theme(legend.title = element_text(size = 25,face = "bold"),
        legend.position = "bottom",
        legend.text = element_text(size = 10),
        strip.text.y.right = element_blank()) +
  labs(
    x = "",
    y = "", # Flights(per day)
    title = "Flights per day of 3 Departure Airport ",
    subtitle = "Line reference average flights per day and  quarter ",
    caption = "Source from nycflights13 package"
  )
```

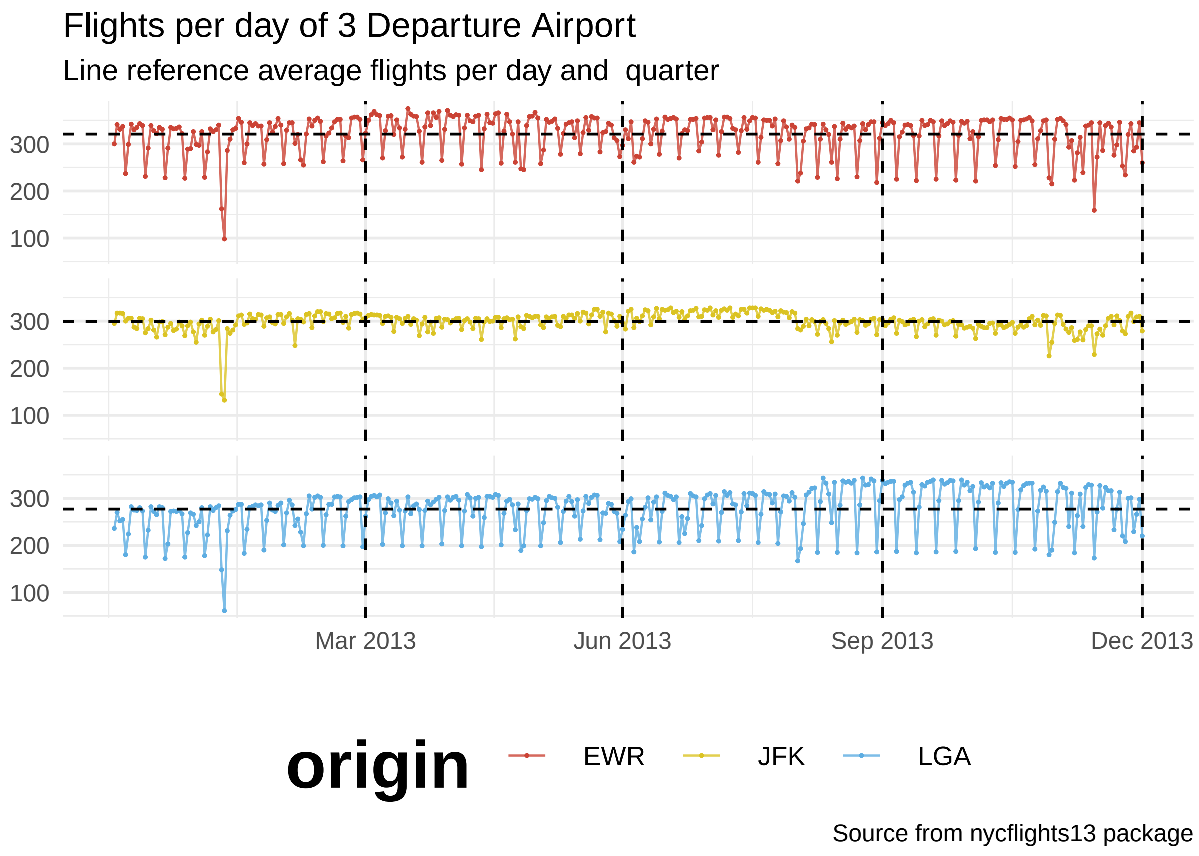


Chart 5 : Pie chart

Departure delays of EWR - Newark Liberty International Airport

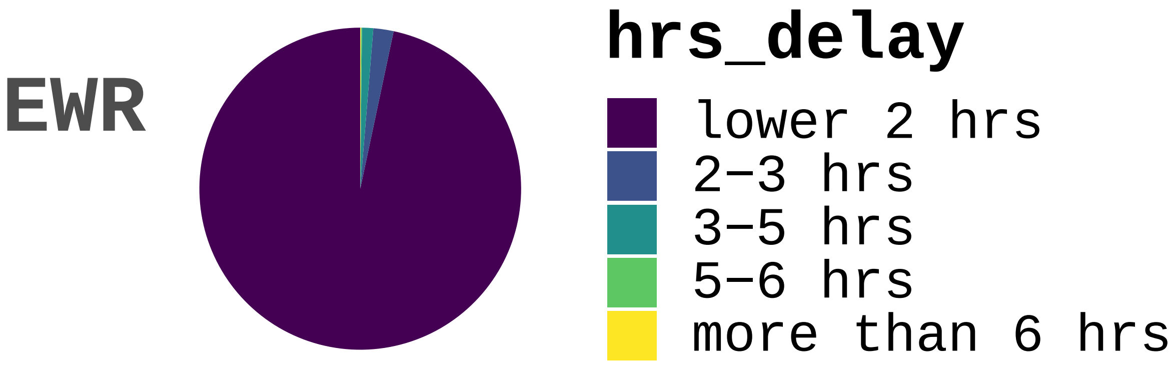
```
##Prepare Data
# No cancelled flights
df_nc <- filter(flights,!is.na(arr_time),!is.na(air_time))

# split departure delays by hours
dt2 <- df_nc %>%
  mutate( hrs_delay =
            case_when(
              dep_delay < 120 ~ "lower 2 hrs", ## <2
              dep_delay < 180 ~ "2-3 hrs", ## <3
              dep_delay < 300 ~ "3-5 hrs", ## <5
              dep_delay < 360 ~ "5-6 hrs", ## <6
              dep_delay >= 360 ~ "more than 6 hrs"
            ))

dt2 <- dt2 %>%
  mutate(hrs_delay = factor(
    hrs_delay,
    levels = c("lower 2 hrs", "2-3 hrs", "3-5 hrs", "5-6 hrs", "more than 6 hrs"),
    labels = c("lower 2 hrs", "2-3 hrs", "3-5 hrs", "5-6 hrs", "more than 6 hrs"),
    ordered = TRUE
  ))

# Find percentage of departure delays of EWR - Newark Liberty International Airport
dt3 <- dt2 %>%
  group_by(hrs_delay,origin) %>%
  count(origin, name = "dep_delay2") %>%
  group_by(origin) %>%
  mutate(percentage = 100*(dep_delay2/sum(dep_delay2))) %>%
  filter(origin == "EWR")
knitr::kable(
  dt2 %>%
  group_by(hrs_delay,origin) %>%
  count(origin, name = "dep_delay2") %>%
  group_by(origin) %>%
  mutate(percentage = dep_delay2/sum(dep_delay2)*100) %>%
  mutate(percentage = round(percentage, digits = 2)) %>%
  mutate(percentage = paste0(percentage," %")) %>%
  filter(origin == "EWR") %>%
  select(origin,hrs_delay,percentage),
  caption = "The percentage of departure delays of EWR - Newark Liberty International Airport"
)
```

Table 6: The percentage of departure delays of EWR - Newark Liberty International Airport

origin	hrs_delay	percentage
EWR	lower 2 hrs	96.67 %
EWR	2-3 hrs	2.03 %
EWR	3-5 hrs	1.12 %

origin	hrs_delay	percentage
EWR	5-6 hrs	0.11 %
EWR	more than 6 hrs	0.07 %

```
## Plot graph by ggplot2
# pie chart
ggplot(data = dt3, mapping = aes(x = origin,
                                 y = percentage,
                                 fill = hrs_delay)) +
  geom_col() +
  scale_y_continuous(breaks = NULL) +
  coord_polar(theta = "y") +  ## pie chart
  theme_classic() +
  theme(legend.title = element_text(size = 25, family = "mono", face = "bold"),
        legend.text = element_text(size = 20, color = "black", family = "mono"),
        legend.key.size = unit(20, "pt"),
        axis.line = element_blank(),  ## out line chart
        axis.ticks = element_blank(), ## out "-"
        axis.text.y = element_text(size = 30, face = "bold", family = "mono")) +
  labs(x = "",
       y = "",
       title = "Departure delays of EWR - Newark Liberty International Airport",
       subtitle = "separate by lower 2 hrs, 2-3 hrs, 3-5 hrs, 5-6 hrs and more than 6 hrs",
       caption = "Source from nycflights13 package")
```



Departure delays of EWR − Newark Liberty International Airport

separate by lower 2 hrs, 2−3 hrs, 3−5 hrs, 5−6 hrs and more than 6 hrs

Source from nycflights13 package

Description : In Thailand, If airline is delayed we are able to claim compensation. The delay level will be divided into hours, starting from 2 hours or more.