HW 01 : NYC Flights 2013 Analysis

Install Package

```
install.packages("nycflights13")

library(tidyverse)
library(nycflights13)
```

```
Updating HTML index of packages in '.Library'

Making 'packages.html' ...
 done

Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts
✗ dplyr::filter()  masks stats::filter()
✗ purrr::flatten() masks jsonlite::flatten()
✗ dplyr::lag()     masks stats::lag()
```

View Data

```
glimpse(flights)
cat("\n")
glimpse(airlines)
cat("\n")
glimpse(airports)
cat("\n")
glimpse(planes)
cat("\n")
glimpse(weather)
```

```
Rows: 336,776
Columns: 19
$ year           <int> 2013, 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, 1, 1
$ day            <int> 1, 1, 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, 558, 55
$ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 60
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2,
$ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 8
$ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 8
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7,
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6"
$ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LG
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IA
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59
```

```
apply(flights, MARGIN = 2, function(col) sum(is.na(col)))
```

```
year:          0 month:          0 day:          0 dep_time:          8255 sched_dep_time:          0 dep_delay:
          8255 arr_time:          8713 sched_arr_time:          0 arr_delay:          9430 carrier:          0 flight:
          0 tailnum:          2512 origin:          0 dest:          0 air_time:          9430 distance:          0 hour:
          0 minute:          0 time_hour:          0
```

Q1: Which airlines had the least departure delay at the beginning of 2013?

```
flights %>%
    filter(year == 2013, month == 1, day == 1, dep_delay >0) %>%
    select(carrier, dep_delay) %>%
    arrange(dep_delay) %>%
    head(10) %>%
    inner_join(airlines, by = "carrier") %>%
    rename( Departure_delayed_time = dep_delay,
            Carrier_name = name)
```

A tibble: 10 × 3

carrier	Departure_delayed_time	Carrier_name
<chr>	<dbl>	<chr>
B6	1	JetBlue Airways
UA	1	United Air Lines Inc.
UA	1	United Air Lines Inc.
UA	1	United Air Lines Inc.
B6	1	JetBlue Airways
B6	1	JetBlue Airways
UA	1	United Air Lines Inc.
B6	1	JetBlue Airways
UA	1	United Air Lines Inc.
VX	1	Virgin America

Q2: Which carrier had the most delay of departure?

```
flights %>%
    filter(dep_delay > 0) %>%
    select(carrier, dep_delay) %>%
    group_by(carrier) %>%
    summarise(dep_delay = sum(dep_delay > 0)) %>%
    arrange(desc(dep_delay)) %>%
    head(5) %>%
    inner_join(airlines, by = "carrier") %>%
    rename( Carrier_ID = carrier,
            Departure_delayed_time = dep_delay,
            Carrier_name = name)
```

A tibble: 5 × 3

Carrier_ID	Departure_delayed_time	Carrier_name
<chr>	<int>	<chr>
UA	27261	United Air Lines Inc.
EV	23139	ExpressJet Airlines Inc.
B6	21445	JetBlue Airways
DL	15241	Delta Air Lines Inc.
AA	10162	American Airlines Inc.

Q3: Which destinations did people travel to the most during the summer (June, July, and August) of 2013?

```
flights %>%
    filter(!is.na(dep_time), year == 2013, month %in% c(6,7,8)) %>%
    select(year,month, dest) %>%
    mutate(seasonal = case_when(month %in% c(6,7,8) ~ "Summer",
                                         month %in% c(9,10,11) ~ "Autumn",
                                         month %in% c(12,1,2) ~ "Winter",
                                         month %in% c(3,4,5) ~ "Spring")) %>%
    group_by(year, seasonal, dest) %>%
    summarise(n_filghts = n()) %>%
    arrange(desc(n_filghts)) %>%
    head(5) %>%
    inner_join(airports %>%
                    select(faa, name),
               by = c("dest" = "faa")) %>%
    rename( Desitination = dest,
            Carrier_name = name)
```

A grouped_df: 5 × 5

year	seasonal	Desitination	n_filghts	Carrier_name
<int>	<chr>	<chr>	<int>	<chr>
2013	Summer	ORD	4528	Chicago Ohare Intl
2013	Summer	LAX	4424	Los Angeles Intl
2013	Summer	ATL	4349	Hartsfield Jackson Atlanta Intl
2013	Summer	BOS	3924	General Edward Lawrence Logan Intl
2013	Summer	SFO	3667	San Francisco Intl

```
`summarise()` has grouped output by 'year', 'seasonal'. You can override usi
```

Q4: In 2013, what season did most people travelled from New York City to other destination?

```
df_seasonals <- mutate(flights,
                    seasonal = case_when(month %in% c(6,7,8) ~ "Summer",
                                         month %in% c(9,10,11) ~ "Autumn",
                                         month %in% c(12,1,2) ~ "Winter",
                                         month %in% c(3,4,5) ~ "Spring"))
df_seasonals %>%
count(seasonal) %>%
arrange(desc(n)) %>%
rename(n_flight = n)
```

A tibble: 4 × 2

seasonal	n_flight
<chr>	<int>
Summer	86995
Spring	85960
Autumn	83731
Winter	80090

Q5: How much was the total traveled distance of each airline each month in 2013?

```
flights %>%
    filter(!is.na(dep_time)) %>%
    inner_join(airlines, by = "carrier") %>%
    group_by(month, carrier, name) %>%
    summarise(n_flights = n(),
              total_distance_in_miles = round(sum(distance), 2)) %>%
    arrange(carrier)
```

A grouped_df: 185 × 5

month	carrier	name	n_flights	total_distance_in_miles
<int>	<chr>	<chr>	<int>	<dbl>
1	9E	Endeavor Air Inc.	1498	717534
2	9E	Endeavor Air Inc.	1353	637366
3	9E	Endeavor Air Inc.	1514	723266
4	9E	Endeavor Air Inc.	1407	691754
5	9E	Endeavor Air Inc.	1388	701809
6	9E	Endeavor Air Inc.	1276	677990
7	9E	Endeavor Air Inc.	1364	736838
8	9E	Endeavor Air Inc.	1378	748201
9	9E	Endeavor Air Inc.	1477	858655
10	9E	Endeavor Air Inc.	1642	976350
11	9E	Endeavor Air Inc.	1575	924766
12	9E	Endeavor Air Inc.	1544	862113
1	AA	American Airlines Inc.	2735	3700495
2	AA	American Airlines Inc.	2405	3250603
3	AA	American Airlines Inc.	2746	3705882
4	AA	American Airlines Inc.	2663	3579654
5	AA	American Airlines Inc.	2770	3714520
6	AA	American Airlines Inc.	2700	3610326
7	AA	American Airlines Inc.	2797	3715661
8	AA	American Airlines Inc.	2830	3754622
9	AA	American Airlines Inc.	2584	3471602
10	AA	American Airlines Inc.	2706	3619786
11	AA	American Airlines Inc.	2558	3443842
12	AA	American Airlines Inc.	2599	3537725
1	AS	Alaska Airlines Inc.	62	148924
2	AS	Alaska Airlines Inc.	54	129708
3	AS	Alaska Airlines Inc.	62	148924
4	AS	Alaska Airlines Inc.	60	144120
5	AS	Alaska Airlines Inc.	62	148924
6	AS	Alaska Airlines Inc.	60	144120
⋮	⋮	⋮	⋮	⋮
7	VX	Virgin America	486	1215475
8	VX	Virgin America	489	1223059
9	VX	Virgin America	453	1132851
10	VX	Virgin America	471	1177783
11	VX	Virgin America	443	1108258
12	VX	Virgin America	468	1169900
1	WN	Southwest Airlines Co.	985	928940
2	WN	Southwest Airlines Co.	861	817643
3	WN	Southwest Airlines Co.	955	940306
4	WN	Southwest Airlines Co.	966	955058
5	WN	Southwest Airlines Co.	988	979018

```
`summarise()` has grouped output by 'month', 'carrier'. You can override usi
```

6	WN	Southwest Airlines Co.	1018	1033205
7	WN	Southwest Airlines Co.	1065	1081410
8	WN	Southwest Airlines Co.	1047	1065845
9	WN	Southwest Airlines Co.	1007	1024462
10	WN	Southwest Airlines Co.	1089	1103138
11	WN	Southwest Airlines Co.	1028	1035873
12	WN	Southwest Airlines Co.	1074	[illegible]
1	YV	Mesa Airlines Inc.	39	8931
2	YV	Mesa Airlines Inc.	46	10534
3	YV	Mesa Airlines Inc.	18	4122
4	YV	Mesa Airlines Inc.	36	14859
5	YV	Mesa Airlines Inc.	44	17951

Q6: What are the fastest flights compared to the distance traveled?

```
flights %>%
  select(carrier, air_time, distance) %>%
  mutate(distance_interval = case_when(
    distance <= 500 ~ "less than 500 miles",
    distance <= 1000 ~ "500 - 1000 miles",
    distance <= 1500 ~ "1000 - 1500 miles",
    TRUE ~ "more than 1500 miles"
  )) %>%
  group_by(carrier, distance_interval) %>%
  summarise(min_air_time = min(air_time, na.rm = T)) %>%
  arrange(min_air_time)
```

A grouped_df: 44 × 3

carrier	distance_interval	min_air_time
<chr>	<chr>	<dbl>
EV	less than 500 miles	20
9E	less than 500 miles	21
US	less than 500 miles	21
UA	less than 500 miles	23
DL	less than 500 miles	26
AA	less than 500 miles	29
B6	less than 500 miles	29
WN	less than 500 miles	31
YV	less than 500 miles	32
MQ	less than 500 miles	33
OO	less than 500 miles	50
FL	less than 500 miles	53
EV	500 - 1000 miles	55
DL	500 - 1000 miles	65
US	500 - 1000 miles	67
MQ	500 - 1000 miles	68
YV	500 - 1000 miles	69
B6	500 - 1000 miles	71
9E	500 - 1000 miles	72
FL	500 - 1000 miles	86
UA	500 - 1000 miles	87
WN	500 - 1000 miles	89
EV	1000 - 1500 miles	93
AA	500 - 1000 miles	94
DL	1000 - 1500 miles	105
B6	1000 - 1500 miles	118
MQ	1000 - 1500 miles	119
UA	1000 - 1500 miles	119
AA	1000 - 1500 miles	125
9E	1000 - 1500 miles	127
OO	500 - 1000 miles	132
WN	1000 - 1500 miles	142
OO	1000 - 1500 miles	152
DL	more than 1500 miles	170
B6	more than 1500 miles	172
AA	more than 1500 miles	173
UA	more than 1500 miles	173
WN	more than 1500 miles	180
F9	more than 1500 miles	195
9E	more than 1500 miles	209
US	more than 1500 miles	243
VX	more than 1500 miles	264

AS	more than 1500 miles	277

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

HW 02 : PostgreSQL Database

```
# 3 dataframes
myFavSeries <- data.frame(
  serie_id = 1:5,
  serie = c("Dark", "Arcane", "Game of Thrones", "Twenty Five Twenty One", "Pe
  score = c(10, 9, 7, 8, 8)
)

genres <- data.frame(
  genre_id = 1:5,
  genre = c("Drama", "Action", "Mystery", "Comedy", "Crime")
)

bridge <- data.frame(
  serie_id = c(1, 1, 2, 3, 3, 4, 4, 5, 5, 5),
  genre_id = c(1, 3, 2, 1, 2, 1, 4, 1, 2, 5)
)
```