NYC Flights 2013 Analysis (R_homework_batch6_data_transformation)

Installation

```
library(tidyverse)
library(dplyr)
```

```
# import flies
df_flights <- read.csv("flights.csv")
df_airlines <- read.csv("airlines.csv")
df_airports <- read.csv("airports.csv")
```

Data View

```
glimpse(df_flights)
glimpse(df_airlines)
glimpse(df_airports)
```

```
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
$ day            <int> 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      <int> 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      <int> 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      <int> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149
$ distance      <int> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73
$ hour          <int> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6
```

Q1: Categorise Data to 4 Quarters (Displaying Overall)

```
overall <- mutate(df_flights, period = case_when(
                    month %in% c(1, 2, 3) ~ "1st Quarter",
                    month %in% c(4, 5, 6) ~ "2nd Quarter",
                    month %in% c(7, 8, 9) ~ "3rd Quarter",
                    month %in% c(10, 11, 12) ~ "4th Quarter"
                ))

overall %>%
    count(period) %>%
    arrange(desc(n))
```

A data.frame: 4 × 2

period	n
<chr>	<int>
3rd Quarter	86326
2nd Quarter	85369
4th Quarter	84292
1st Quarter	80789

Q2: Top 5 of Popular Destinations (Arrival Flights) During The First Quarter

```
#first quarter (Jan-Apr)
pop_des <- df_flights %>%
                filter(month <= 3) %>%
                select(dest) %>%
                group_by(dest) %>%
                summarise(n = n()) %>%
```

```
                arrange(desc(n)) %>%
                head(5)

# align table
port_names <- df_airports %>% select(faa, name)
pop_des %>% merge(pop_des = "dest", port_names, by = 1) %>%
            select(1, 3, 2) %>%
            arrange(desc(n)) %>%
            rename(FAA = dest, Total_flights = n, Airport_names = name)
```

A data.frame: 5 × 3

FAA	Airport_names	Total_flights
<chr>	<chr>	<int>
ATL	Hartsfield Jackson Atlanta Intl	4111
ORD	Chicago Ohare Intl	3809
BOS	General Edward Lawrence Logan Intl	3751
MCO	Orlando Intl	3550
FLL	Fort Lauderdale Hollywood Intl	3472

Q3: Top 10 of Bad Travelling Experiences Provided Airlines (Time Delay)

```
delay <- df_flights %>%
            mutate(df_flights,
                    int_dep_delay = abs(dep_delay),
                    int_arr_delay = abs(arr_delay)) %>%
                    select(carrier, int_dep_delay, int_arr_delay) %>%
                    mutate(total_delay = int_dep_delay + int_arr_delay)
delay %>%
        group_by(carrier) %>%
        summarise(total_delay=sum(is.na(total_delay))) %>%
        left_join(df_airlines, by = "carrier") %>%
        select(name, total_delay) %>%
        arrange(desc(total_delay)) %>%
        rename(Airlines = name, Total_delay.mins = total_delay) %>%
        head(10)
```

A tibble: 10 × 2

Airlines	Total_delay.mins
<chr>	<int>
ExpressJet Airlines Inc.	3065
Envoy Air	1360
Endeavor Air Inc.	1166
United Air Lines Inc.	883
American Airlines Inc.	782
US Airways Inc.	705
JetBlue Airways	586
Delta Air Lines Inc.	452
Southwest Airlines Co.	231
AirTran Airways Corporation	85

Q4: Overall Average Flying Speed of Each Airline

```
# formular speed = distance /(air_time/60) (miles per hour)

# clean missing values

clean_flights <- drop_na(df_flights)
clean_flights %>% head(10)

# all flights
fly_spd <- clean_flights %>%
            select(carrier, air_time, distance) %>%
            mutate(flying_spd = distance/(air_time/60)) %>%
            tibble()

# group_by airlines
fly_spd %>% group_by(carrier) %>%
            summarise(avg_fly_spd = mean(flying_spd)) %>%
            left_join(df_airlines, by = "carrier") %>%
            select(name, avg_fly_spd) %>%
            arrange(desc(avg_fly_spd)) %>%
            rename(Airlines = name, Avg_mph = avg_fly_spd)
```

A data.frame: 10 × 19

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	f
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<
1	2013	1	1	517	515	2	830	819	11	UA	1
2	2013	1	1	533	529	4	850	830	20	UA	1
3	2013	1	1	542	540	2	923	850	33	AA	1
4	2013	1	1	544	545	-1	1004	1022	-18	B6	7
5	2013	1	1	554	600	-6	812	837	-25	DL	4
6	2013	1	1	554	558	-4	740	728	12	UA	1
7	2013	1	1	555	600	-5	913	854	19	B6	5
8	2013	1	1	557	600	-3	709	723	-14	EV	5
9	2013	1	1	557	600	-3	838	846	-8	B6	7
10	2013	1	1	558	600	-2	753	745	8	AA	3

A tibble: 16 × 2

Airlines	Avg_mph
<chr>	<dbl>
Hawaiian Airlines Inc.	480.3577
Virgin America	446.1749
Alaska Airlines Inc.	443.6789
Frontier Airlines Inc.	425.1721
United Air Lines Inc.	420.8838
Delta Air Lines Inc.	418.4628
American Airlines Inc.	417.4727
Southwest Airlines Co.	400.5320
JetBlue Airways	399.9715
AirTran Airways Corporation	394.3581
Envoy Air	368.4028
SkyWest Airlines Inc.	366.3201
ExpressJet Airlines Inc.	362.9436
Endeavor Air Inc.	345.4304
US Airways Inc.	341.9397
Mesa Airlines Inc.	331.9700

Q5: In The 3rd Quarter, Which Airports Have The Most Departure Flights

```
dep_flights <- df_flights %>%
                filter(month >=  7 & month <= 9) %>%
                select(origin) %>%
                group_by(origin) %>%
                summarise(n = n()) %>%
                arrange(desc(n))

# align table
port_names <- df_airports %>% select(faa, name)
dep_flights %>% merge(dep_flights = "origin", port_names, by = 1) %>%
            select(3, 2) %>%
            arrange(desc(n)) %>%
            rename(Total_flights = n, Airports = name)
```

A data.frame: 3 × 2

Airports	Total_flights
<chr>	<int>
Newark Liberty Intl	30384
John F Kennedy Intl	28914
La Guardia	27028