NYC flights 2013 Analysis

```
install.packages("nycflights13")
library(nycflights13)
library(tidyverse)
library(dplyr)
```

```
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 1.3.1

✓ 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()
```

```
data(package="nycflights13")
```

Data sets

A data.frame: 5 × 3

Package	Item	Title
<chr>	<chr>	<chr>
nycflights13	airlines	Airline names.
nycflights13	airports	Airport metadata
nycflights13	flights	Flights data
nycflights13	planes	Plane metadata.
nycflights13	weather	Hourly weather data

```
glimpse(flights)
```

```
Rows: 336,776
Columns: 19
$ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2
$ month          <int> 1, 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, 1
$ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, 
$ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, 
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1
$ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,
$ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "
$ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, 
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0
```

```
glimpse(airlines)
```

```
Rows: 16
Columns: 2
$ carrier <chr> "9E", "AA", "AS", "B6", "DL", "EV", "F9", "FL", "HA", "MQ", "O…
$ name    <chr> "Endeavor Air Inc.", "American Airlines Inc.", "Alaska Airline…
```

```
glimpse(airports)
```

```
Rows: 1,458
Columns: 8
$ faa   <chr> "04G", "06A", "06C", "06N", "09J", "0A9", "0G6", "0G7", "0P2", "…
$ name  <chr> "Lansdowne Airport", "Moton Field Municipal Airport", "Schaumbur…
$ lat   <dbl> 41.13047, 32.46057, 41.98934, 41.43191, 31.07447, 36.37122, 41.4…
$ lon   <dbl> -80.61958, -85.68003, -88.10124, -74.39156, -81.42778, -82.17342…
$ alt   <dbl> 1044, 264, 801, 523, 11, 1593, 730, 492, 1000, 108, 409, 875, 10…
$ tz    <dbl> -5, -6, -6, -5, -5, -5, -5, -5, -5, -8, -5, -6, -5, -5, -5, -5, …
$ dst   <chr> "A", "A", "A", "A", "A", "A", "A", "A", "U", "A", "A", "U", "A",…
$ tzone <chr> "America/New_York", "America/Chicago", "America/Chicago", "Ameri…
```

```
glimpse(planes)
```

```
Rows: 3,322
Columns: 9
$ tailnum      <chr> "N10156", "N102UW", "N103US", "N104UW", "N10575", "N105UW…
$ year         <int> 2004, 1998, 1999, 1999, 2002, 1999, 1999, 1999, 1999, 199…
$ type         <chr> "Fixed wing multi engine", "Fixed wing multi engine", "Fi…
$ manufacturer <chr> "EMBRAER", "AIRBUS INDUSTRIE", "AIRBUS INDUSTRIE", "AIRBU…
$ model        <chr> "EMB-145XR", "A320-214", "A320-214", "A320-214", "EMB-145…
$ engines      <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
$ seats        <int> 55, 182, 182, 182, 55, 182, 182, 182, 182, 182, 55, 55, 5…
$ speed        <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
$ engine       <chr> "Turbo-fan", "Turbo-fan", "Turbo-fan", "Turbo-fan", "Turb…
```

```
glimpse(weather)
```

```
Rows: 26,115
Columns: 15
$ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EW…
$ year       <int> 2013, 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, 1, 1,…
$ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ hour       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, …
$ temp       <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39.92, 39.…
$ dewp       <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28.04, 28.…
$ humid      <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62.21, 62.…
$ wind_dir   <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 260, 330,…
$ wind_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11.50780, …
$ wind_gust  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20.…
$ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ pressure   <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 1012.2, 101…
$ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,…
$ time_hour  <dttm> 2013-01-01 06:00:00, 2013-01-01 07:00:00, 2013-01-01 08:00…
```

```
#example : which carrier had most flights in May 2013
flights %>%
  filter(month == 5, year == 2013) %>%
  count(carrier) %>%
  arrange(desc(n)) %>%
  left_join(airlines, by = "carrier") %>%
  head(5)
```

A tibble: 5 × 3

carrier	n	name
<chr>	<int>	<chr>
UA	4960	United Air Lines Inc.
EV	4817	ExpressJet Airlines Inc.
B6	4576	JetBlue Airways
DL	4082	Delta Air Lines Inc.
AA	2803	American Airlines Inc.

```
weather %>%
    drop_na(temp) %>%
    select(origin,temp) %>%
    mutate(temp_group =
            if_else (temp >= 33, "low temp", "high temp")) %>%
    group_by(origin) %>%
    count(temp_group) %>%
    arrange(desc(n))
```

A grouped_df: 6 × 3

origin	temp_group	n
<chr>	<chr>	<int>
LGA	low temp	7828
JFK	low temp	7782
EWR	low temp	7660
EWR	high temp	1042
JFK	high temp	924
LGA	high temp	878

```
## 01 which carrier had most destinattion = ORD or ATL
flights %>%
  filter(dest %in% c("ORD","ATL")) %>%
  count(carrier) %>%
  arrange(desc(n)) %>%
  rename(count_dest = n) %>%
  left_join(airlines, by = "carrier") %>%
  head(5)
```

A tibble: 5 × 3

carrier	count_dest	name
<chr>	<int>	<chr>
DL	10571	Delta Air Lines Inc.
UA	7087	United Air Lines Inc.
AA	6059	American Airlines Inc.
MQ	4598	Envoy Air
FL	2337	AirTran Airways Corporation

```
# Q2 Top 5 routes (origin -> dest)
flights %>%
    filter(!is.na(dep_time) & !is.na(arr_time)) %>%
    group_by(origin, dest) %>%
    count(dest) %>%
    arrange(desc(n)) %>%
    head(5)
```

A grouped_df: 5 × 3

origin	dest	n
<chr>	<chr>	<int>
JFK	LAX	11182
LGA	ATL	10063
LGA	ORD	8529
JFK	SFO	8126
LGA	CLT	5963

```
# 03 Which month has the highest average temperature?
weather %>%
    select (month, temp) %>%
    filter(!is.na(temp)) %>%
    group_by(month) %>%
    summarize(mean_temp = mean(temp)) %>%
    arrange(desc(mean_temp)) %>%
    head(1)
```

A tibble: 1 × 2

month	mean_temp
<int>	<dbl>
7	80.06622

```
# 04 Top 5 of airline has the most % arrival delays
delay <- flights %>%
    filter(arr_delay > 0) %>%
    count(carrier) %>%
    rename(count_delay = n)
flight <- flights %>%
    count(carrier) %>%
    rename(count_flight= n) %>%
    left_join(airlines, by = "carrier")
percent <- flight %>%
    left_join(delay, by = "carrier") %>%
    mutate(percent_delay = (count_delay/count_flight)*100 ) %>%
    arrange(desc(percent_delay)) %>%
    select(name,carrier,percent_delay) %>%
    head(5)
percent
```

A tibble: 5 × 3

name	carrier	percent_delay
<chr>	<chr>	<dbl>
AirTran Airways Corporation	FL	58.12883
Frontier Airlines Inc.	F9	57.22628
ExpressJet Airlines Inc.	EV	45.19595
Envoy Air	MQ	44.29670
JetBlue Airways	B6	43.21223

```
# 05 How many flights have the longest distance?
# Calculated with complete 'dep_time' and 'arr_time' data only
flights %>%
    filter(!is.na(dep_time) & !is.na(arr_time)) %>%
    group_by(origin, dest, distance) %>%
    count(distance) %>%
    arrange(desc(distance)) %>%
    rename(number_flights = n,
    distance_mile = distance) %>%

    # converting mile to kilometre values
    mutate(distance_km = distance_mile * 1.609) %>%
    select(origin, dest, distance_mile, distance_km, number_flights) %>%
    head(5)
```

A grouped_df: 5 × 5

origin	dest	distance_mile	distance_km	number_flights
<chr>	<chr>	<dbl>	<dbl>	<int>
JFK	HNL	4983	8017.647	342
EWR	HNL	4963	7985.467	363
EWR	ANC	3370	5422.330	8
JFK	SFO	2586	4160.874	8126
JFK	OAK	2576	4144.784	311