```
library(dplyr)

library(tidyverse)
```

```
flights <- read.csv("flights.csv", stringsAsFactors = FALSE)

glimpse(flights)

flights %>% distinct(origin)
```

```
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
$ minute         <int> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59
```

A data.frame: 3 × 1

origin
<chr>
EWR
LGA
JFK

```
airlines <- read.csv("airlines.csv")

glimpse(airlines)
```

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

1. Top 10 longest flights in 2013

```
top10 <- flights %>%
    select(1:3,10,13:15) %>%
    arrange(desc(air_time)) %>%
    head(10)

top10 %>%
    left_join(airlines, by = "carrier") %>%
    select(1,2,3, carrier_name = name, 5:7)
```

A data.frame: 10 × 7

year	month	day	carrier_name	origin	dest	air_time
<int>	<int>	<int>	<chr>	<chr>	<chr>	<int>
2013	3	17	United Air Lines Inc.	EWR	HNL	695
2013	2	6	Hawaiian Airlines Inc.	JFK	HNL	691
2013	3	15	Hawaiian Airlines Inc.	JFK	HNL	686
2013	3	17	Hawaiian Airlines Inc.	JFK	HNL	686
2013	3	16	Hawaiian Airlines Inc.	JFK	HNL	683
2013	2	5	Hawaiian Airlines Inc.	JFK	HNL	679
2013	11	12	United Air Lines Inc.	EWR	HNL	676
2013	3	14	Hawaiian Airlines Inc.	JFK	HNL	676
2013	11	20	Hawaiian Airlines Inc.	JFK	HNL	675
2013	3	15	United Air Lines Inc.	EWR	HNL	671

2. The most delayed airline from LGA in August 2013

```
flights %>%
    select(1,2,6,10) %>%
    filter(year == 2013 & month == 8 & dep_delay > 0) %>%
    left_join(airlines, by = "carrier") %>%
    group_by(name) %>%
    summarise(n = n()) %>%
    arrange(desc(n))
```

A tibble: 16 × 2

name	n
<chr>	<int>
United Air Lines Inc.	2518
JetBlue Airways	2179
ExpressJet Airlines Inc.	1743
Delta Air Lines Inc.	1547
American Airlines Inc.	841
Envoy Air	716
Southwest Airlines Co.	675
Endeavor Air Inc.	639
US Airways Inc.	442
Virgin America	162
AirTran Airways Corporation	159
Frontier Airlines Inc.	46
Mesa Airlines Inc.	20
Alaska Airlines Inc.	17
Hawaiian Airlines Inc.	6
SkyWest Airlines Inc.	3

3. Average flight duration in from JFK to HNL in July 2013 (group by airlines)

```
flights %>%
    select(1,2,10,13:15) %>%
    filter(origin == "JFK" & dest == "HNL") %>%
    left_join(airlines, by = "carrier") %>%
    group_by(name) %>%
    summarise(average_time_spent = mean(air_time))
```

A tibble: 1 × 2

name	average_time_spent
<chr>	<dbl>
Hawaiian Airlines Inc.	623.0877

4. Top 10 distance flight of United airlines

```
flights %>%
    arrange(desc(distance)) %>%
    filter(carrier == "UA") %>%
    select(16,13,14) %>%
    distinct(distance,origin,dest) %>%
    head(10)
```

A data.frame: 10 × 3

	distance	origin	dest
	<int>	<chr>	<chr>
1	4963	EWR	HNL
2	3370	EWR	ANC
3	2586	JFK	SFO
4	2565	EWR	SFO
5	2475	JFK	LAX
6	2454	EWR	LAX
7	2434	EWR	SNA
8	2434	EWR	PDX
9	2425	EWR	SAN
10	2402	EWR	SEA

5. the most flight departure to EWR in 2013 by month

```
month <- c(1:12)
month_name <- c("Jan","Feb","Mar",
                "Apr","May","Jun",
                "Jul","Aug","Sep",
                "Oct","Nov","Dec")

my_month <- data.frame(month, month_name)


flights %>%
    filter(origin == "EWR") %>%
    group_by(month) %>%
    summarise(num_departure = n()) %>%
    arrange(desc(num_departure)) %>%
    inner_join(my_month, by = "month") %>%
    select( month = month_name,
            num_departure)
```

A tibble: 12 × 2

month	num_departure
<chr>	<int>
May	10592
Apr	10531
Jul	10475
Mar	10420
Aug	10359
Jun	10175
Oct	10104
Dec	9922
Jan	9893
Nov	9707
Sep	9550
Feb	9107