Install Packages

```
install.packages("nycflights13")
```

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

Making 'packages.html' ...
 done
```

Load Library

```
library(nycflights13)
library(tidyverse)
library(dplyr)
```

```
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()
```

Preview Dataset

```
# check what datasets are in packages "nycflights13" ?
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

```
mean(complete.cases(airlines))
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…
```

1

```
glimpse(airports)
mean(complete.cases(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…
```

0.997942386831276

```
clean_airports <- airports[complete.cases(airports[,]),]
clean_airports
mean(complete.cases(clean_airports))
```

A tibble: 1455 × 8

faa	name	lat	lon	alt	tz	dst	tzone
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>
04G	Lansdowne Airport	41.13047	-80.61958	1044	-5	A	America/New_York
06A	Moton Field Municipal Airport	32.46057	-85.68003	264	-6	A	America/Chicago
06C	Schaumburg Regional	41.98934	-88.10124	801	-6	A	America/Chicago
06N	Randall Airport	41.43191	-74.39156	523	-5	A	America/New_York
09J	Jekyll Island Airport	31.07447	-81.42778	11	-5	A	America/New_York
0A9	Elizabethton Municipal Airport	36.37122	-82.17342	1593	-5	A	America/New_York
0G6	Williams County Airport	41.46731	-84.50678	730	-5	A	America/New_York
0G7	Finger Lakes Regional Airport	42.88356	-76.78123	492	-5	A	America/New_York
0P2	Shoestring Aviation Airfield	39.79482	-76.64719	1000	-5	U	America/New_York
0S9	Jefferson County Intl	48.05381	-122.81064	108	-8	A	America/Los_Angeles
0W3	Harford County Airport	39.56684	-76.20240	409	-5	A	America/New_York
10C	Galt Field Airport	42.40289	-88.37511	875	-6	U	America/Chicago
17G	Port Bucyrus-Crawford County Airport	40.78156	-82.97481	1003	-5	A	America/New_York
19A	Jackson County Airport	34.17586	-83.56160	951	-5	U	America/New_York
1A3	Martin Campbell Field Airport	35.01581	-84.34683	1789	-5	A	America/New_York
1B9	Mansfield Municipal	42.00013	-71.19677	122	-5	A	America/New_York
1C9	Frazier Lake Airpark	54.01333	-124.76833	152	-8	A	America/Vancouver
1CS	Clow International Airport	41.69597	-88.12923	670	-6	U	America/Chicago
1G3	Kent State Airport	41.15139	-81.41511	1134	-5	A	America/New_York
1G4	Grand Canyon West Airport	35.89990	-113.81567	4813	-7	A	America/Phoenix
1H2	Effingham Memorial Airport	39.07000	-88.53400	585	-6	A	America/Chicago
1OH	Fortman Airport	40.55533	-84.38662	885	-5	U	America/New_York
1RL	Point Roberts Airpark	48.97972	-123.07889	10	-8	A	America/Los_Angeles
23M	Clarke CO	32.05170	-88.44340	320	-6	A	America/Chicago
24C	Lowell City Airport	42.95392	-85.34391	681	-5	A	America/New_York
24J	Suwannee County Airport	30.30013	-83.02469	104	-5	A	America/New_York
25D	Forest Lake Airport	45.24775	-92.99439	925	-6	A	America/Chicago
29D	Grove City Airport	41.14603	-80.16775	1371	-5	A	America/New_York
2A0	Mark Anton Airport	35.48625	-84.93108	718	-5	A	America/New_York
2B2	Plum Island Airport	42.79536	-70.83944	11	-5	A	America/New_York
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X49	South Lakeland Airport	27.93336	-82.04397	110	-5	A	America/New_York
X59	Valkaria Municipal	27.96086	-80.55833	26	-5	A	America/New_York
XFL	Flagler County Airport	29.28210	-81.12120	33	-5	A	America/New_York
XNA	NW Arkansas Regional	36.28187	-94.30681	1287	-6	A	America/Chicago
XZK	Amherst Amtrak Station AMM	42.37500	-72.51139	258	-5	A	America/New_York
Y51	Municipal Airport	43.57936	-90.89647	1292	-6	A	America/Chicago
Y72	Bloyer Field	43.97622	-90.48061	966	-6	A	America/Chicago
YIP	Willow Run	42.23793	-83.53041	716	-5	A	America/New_York
YKM	Yakima Air Terminal McAllister Field	46.56820	-120.54400	1095	-8	A	America/Los_Angeles
YKN	Chan Gurney	42.87110	-97.39690	1200	-6	A	America/Chicago

YNG	Youngstown Warren Rgnl	41.26074	-80.67910	1196	-5	A	America/New_York
YUM	Yuma Mcas Yuma Intl	32.65658	-114.60598	216	-7	N	America/Phoenix
Z84	Clear	64.30120	-149.12014	552	-9	A	America/Anchorage
ZBP	Penn Station	39.30722	-76.61556	66	-5	A	America/New_York
ZFV	Philadelphia 30th St Station	39.95570	-75.18200	0	-5	A	America/New_York
ZPH	Municipal Airport	28.22806	-82.15583	90	-5	A	America/New_York
ZRA	Atlantic City Rail Terminal	39.36650	-74.44200	8	-5	A	America/New_York
ZRD	Train Station	37.53430	-77.42945	26	-5	A	America/New_York
ZRP	Newark Penn Station	40.73472	-74.16417	0	-5	A	America/New_York
ZRT	Hartford Union Station	41.76888	-72.68150	0	-5	A	America/New_York
ZRZ	New Carrollton Rail Station	38.94800	-76.87190	39	-5	A	America/New_York
ZSF	Springfield Amtrak Station	42.10600	-72.59305	65	-5	A	America/New_York
ZSY	Scottsdale Airport	33.62289	-111.91053	1519	-7	A	America/Phoenix
ZTF	Stamford Amtrak Station	41.04694	-73.54149	0	-5	A	America/New_York
ZTY	Boston Back Bay Station	42.34780	-71.07500	20	-5	A	America/New_York
ZUN	Black Rock	35.08323	-108.79178	6454	-7	A	America/Denver
ZVE	New Haven Rail Station	41.29867	-72.92599	7	-5	A	America/New_York
ZWI	Wilmington Amtrak Station	39.73667	-75.55167	0	-5	A	America/New_York
ZWU	Washington Union Station	38.89746	-77.00643	76	-5	A	America/New_York
ZYP	Penn Station	40.75050	-73.99350	35	-5	A	America/New_York

1

```
glimpse(flights)
mean(complete.cases(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
```

0.971999192341497

```
clean_flights <- flights[complete.cases(flights[,]),]
mean(complete.cases(clean_flights))
```

1

```
glimpse(planes)
mean(complete.cases(planes))
# cannot drop NA from planes dataset because there too many missing values
```

```
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…
```

0.00692354003612282

```
glimpse(weather)
mean(complete.cases(weather))
# cannot drop NA from weather dataset because there too many missing values
```

```
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, 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…
```

0.190695002871913

NYCFLIGHTS13 Analysis

1. In August 2013, Delayed departure / arrivals

1.1 Which top 5 flights is the most delayed departure and which carrier ?

```
q1_1 <- clean_flights %>%
  filter(month == 8, dep_delay > 0) %>%
  arrange(desc(dep_delay)) %>%
  left_join(airlines, by = "carrier") %>%
  select(flight,carrier,name,dep_delay) %>%
  rename(depdelayed_min = dep_delay) %>%
  mutate(depdelayed_hr = round((depdelayed_min/60),2)) %>%
  head(5)

View(q1_1)
```

A tibble: 5 × 5

flight	carrier	name	depdelayed_min	depdelayed_hr
<int>	<chr>	<chr>	<dbl>	<dbl>
4978	EV	ExpressJet Airlines Inc.	520	8.67
843	DL	Delta Air Lines Inc.	508	8.47
1373	DL	Delta Air Lines Inc.	436	7.27
708	UA	United Air Lines Inc.	424	7.07
227	B6	JetBlue Airways	411	6.85

1.2 Which top 5 flights is the most delayed arrivals and which carrier ?

```
q1_2 <- clean_flights %>%
  filter(month == 8, arr_delay > 0) %>%
  arrange(desc(arr_delay)) %>%
  left_join(airlines, by = "carrier") %>%
  select(flight,carrier,name,arr_delay) %>%
  rename(arrdelayed_min = arr_delay) %>%
  mutate(arrdelayed_hr = round((arrdelayed_min/60),2)) %>%
  head(5)

View(q1_2)
```

A tibble: 5 × 5

flight	carrier	name	arrdelayed_min	arrdelayed_hr
<int>	<chr>	<chr>	<dbl>	<dbl>
4978	EV	ExpressJet Airlines Inc.	490	8.17
843	DL	Delta Air Lines Inc.	469	7.82
3074	WN	Southwest Airlines Co.	410	6.83
411	VX	Virgin America	404	6.73
227	B6	JetBlue Airways	403	6.72

2. What is top 3 newest type of engine plane, where are they made from

and how much lifetime are they ?

```
q2 <- planes %>%
  group_by(type) %>%
  arrange(desc(year)) %>%
  slice(1) %>%
  select(type,year,manufacturer) %>%
  mutate(lifetime_yr = 2022-year) %>%
  arrange(lifetime_yr) %>%
  head(3)

View(q2)
```

A grouped_df: 3 × 4

type	year	manufacturer	lifetime_yr
<chr>	<int>	<chr>	<dbl>
Fixed wing multi engine	2013	AIRBUS	9
Rotorcraft	2012	ROBINSON HELICOPTER CO	10
Fixed wing single engine	2007	AVIAT AIRCRAFT INC	15

3. What is the most departure airport in May 2013 ?

```
q3 <- flights %>%
  filter(month == 5) %>%
  group_by(origin) %>%
  count(origin) %>%
  arrange(desc(n)) %>%
  left_join(airports, by = c("origin" = "faa")) %>%
  select(origin,name,n) %>%
  head(1)

View(q3)
```

A grouped_df: 1 × 3

origin	name	n
<chr>	<chr>	<int>
EWR	Newark Liberty Intl	10592

4. What is the shortest flight of each airline ?

```
q4 <- flights %>%
  select(carrier,air_time) %>%
  group_by(carrier) %>%
  arrange(air_time) %>%
  slice(1) %>%
  left_join(airlines, by = "carrier") %>%
  arrange(air_time)

View(q4)
```

A grouped_df: 16 × 3

carrier	air_time	name
<chr>	<dbl>	<chr>
EV	20	ExpressJet Airlines Inc.
9E	21	Endeavor Air Inc.
US	21	US Airways Inc.
UA	23	United Air Lines Inc.
DL	26	Delta Air Lines Inc.
AA	29	American Airlines Inc.
B6	29	JetBlue Airways
WN	31	Southwest Airlines Co.
YV	32	Mesa Airlines Inc.
MQ	33	Envoy Air
OO	50	SkyWest Airlines Inc.
FL	53	AirTran Airways Corporation
F9	195	Frontier Airlines Inc.
VX	264	Virgin America
AS	277	Alaska Airlines Inc.
HA	580	Hawaiian Airlines Inc.

5. How many flights in each month ?

```
q5 <- flights %>%
  group_by(month) %>%
  summarise(n_flight = n()) %>%
  mutate(perc = round((100*n_flight)/sum(n_flight),digits = 2)) %>%
  arrange(month)

View(q5)
```

A tibble: 12 × 3

month	n_flight	perc
<int>	<int>	<dbl>
1	27004	8.02
2	24951	7.41
3	28834	8.56
4	28330	8.41
5	28796	8.55
6	28243	8.39
7	29425	8.74
8	29327	8.71
9	27574	8.19
10	28889	8.58
11	27268	8.10
12	28135	8.35