Import Data

```
# import
library(tidyverse)
flights <- read_csv("flights.csv")
df <- flights
```

```
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()

Rows: 336776 Columns: 19

── Column specification ────────────────────────────────────────────────────
Delimiter: ","
```

Preview Data

```
# preview data
glimpse (flights)
flights %>% head(5)
flights %>% tail(5)
check_na <- function(col) {
    sum(is.na(col))
}
apply(flights, MARGIN = 2, FUN = check_na)
```

```
Rows: 336,776
Columns: 19
$ year           <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2
$ month          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ day            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ dep_time       <dbl> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558,
$ sched_dep_time <dbl> 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       <dbl> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,
$ sched_arr_time <dbl> 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         <dbl> 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
```

A tibble: 5 × 19

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
2013	1	1	517	515	2	830	819	11	UA	1545
2013	1	1	533	529	4	850	830	20	UA	1714
2013	1	1	542	540	2	923	850	33	AA	1141
2013	1	1	544	545	-1	1004	1022	-18	B6	725
2013	1	1	554	600	-6	812	837	-25	DL	461

A tibble: 5 × 19

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
2013	9	30	NA	1455	NA	NA	1634	NA	9E	3393
2013	9	30	NA	2200	NA	NA	2312	NA	9E	3525
2013	9	30	NA	1210	NA	NA	1330	NA	MQ	3461
2013	9	30	NA	1159	NA	NA	1344	NA	MQ	3572
2013	9	30	NA	840	NA	NA	1020	NA	MQ	3531

```
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
```

Data Cleaning

```
# drop na
df <- drop_na(flights)
# glimpse new df
glimpse(df)
apply(df, MARGIN = 2, FUN = check_na)
```

```
Rows: 327,346
Columns: 19
$ year           <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2
$ month          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ day            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
$ dep_time       <dbl> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558,
$ sched_dep_time <dbl> 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       <dbl> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,
$ sched_arr_time <dbl> 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         <dbl> 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
```

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

```
# prepare dep_time, sched_dep_time, dep_delay, arr_time, sched_arr_time, arr_dela
df <- df %>%
    mutate(dep_time_mins = (dep_time %/% 100)*60 + (dep_time %% 100),
           sched_dep_time_mins = (sched_dep_time %/% 100)*60 + (sched_dep_time %%
           arr_time_mins = (arr_time %/% 100)*60 + (arr_time %% 100),
           sched_arr_time_mins = (sched_arr_time %/% 100)*60 + (sched_arr_time %%
           air_time = arr_time_mins - dep_time_mins)

glimpse(df)
df %>% head(5)
df %>% tail(5)
```

```
Rows: 327,346
Columns: 23
$ year                <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20
$ month               <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ day                 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ dep_time            <dbl> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558,
$ sched_dep_time      <dbl> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600,
$ dep_delay           <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -
$ arr_time            <dbl> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753,
$ sched_arr_time      <dbl> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745,
$ arr_delay           <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3,
$ carrier             <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B
$ flight              <dbl> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 3
$ tailnum             <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN",
$ origin              <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "
$ dest                <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "
$ air_time            <dbl> 193, 197, 221, 260, 138, 106, 198, 72, 161, 115, 1
$ distance            <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944,
$ hour                <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6,
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	⋯
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	⋯
2013	1	1	517	515	2	830	819	11	UA	⋯
2013	1	1	533	529	4	850	830	20	UA	⋯
2013	1	1	542	540	2	923	850	33	AA	⋯
2013	1	1	544	545	-1	1004	1022	-18	B6	⋯
2013	1	1	554	600	-6	812	837	-25	DL	⋯

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	⋯
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	⋯
2013	9	30	2240	2245	-5	2334	2351	-17	B6	⋯
2013	9	30	2240	2250	-10	2347	7	-20	B6	⋯
2013	9	30	2241	2246	-5	2345	1	-16	B6	⋯
2013	9	30	2307	2255	12	2359	2358	1	B6	⋯
2013	9	30	2349	2359	-10	325	350	-25	B6	⋯

Q1 3 อันดับเดือนที่มีจำนวนเที่ยวบินมากที่สุด

```
# method 1: count()
df %>%
    count (month) %>%
    arrange (desc(n)) %>%
    head (3)
```

A tibble: 3 × 2

month	n
<dbl>	<int>
8	28756
10	28618
7	28293

```
# method 2: group by + summarise
df %>%
    group_by (month) %>%
    summarise (n = n()) %>%
    arrange (desc(n)) %>%
    head (3)
```

A tibble: 3 × 2

month	n
<dbl>	<int>
8	28756
10	28618
7	28293

Q2 สำรวจสายการบินที่มีจำนวนครั้งการดีเลย์น้อยที่สุด 5 อันดับแรก

```
df %>%
    filter (arr_delay > 0 | dep_delay > 0) %>% # ทั้งออกช้า และ ถึงช้า
    group_by (carrier) %>%
    summarise (n = n()) %>%
    arrange (n) %>%
    head(5)
```

A tibble: 5 × 2

carrier	n
<chr>	<int>
OO	11
HA	129
AS	289
YV	292
F9	476

Q3 สำรวจสถานีที่ผู้คนเข้า-ออกในช่วงเดือน 11 และ 12 (เทศกาลของชาวคริสต์ เช่น ขอบคุณพระเจ้า และคริสต์มาส รวมถึงปีใหม่)

```
# เที่ยวบินขาออก
df %>%
    filter (month == 11 | month == 12) %>%
    group_by (origin) %>%
    summarise (ori_n = n()) %>%
    arrange (desc (ori_n)) %>%


# เที่ยวบินขาเข้า
df %>%
    filter(month == 11 | month == 12) %>%
    group_by (dest) %>%
    summarise (dest_n = n()) %>%
    arrange(desc (dest_n)) %>%
```

A tibble: 3 × 2

origin	ori_n
<chr>	<int>
EWR	19013
JFK	17568
LGA	17410

A tibble: 97 × 2

dest	dest_n
<chr>	<int>
ATL	2807
LAX	2697
ORD	2494
MCO	2354
CLT	2330
SFO	2317
BOS	2314
MIA	2063
FLL	2046
DTW	1424
DCA	1336
TPA	1331
PBI	1302
DFW	1273
DEN	1201
IAH	1198
RDU	1177
MSP	1173
BNA	1084
SJU	972
LAS	896
IAD	829
PHX	790
BUF	724
STL	723
MSY	671
MDW	670
CLE	668
RSW	655
CVG	582
⋮	⋮
BGR	107
TYS	106

GRR	98
STT	92
ALB	90
DSM	90
MHT	74
ABQ	61
BUR	61
PVD	33
EGE	25
JAC	13
EYW	10
BZN	4
SBN	3
HDN	2
MTJ	2
PSP	2
LEX	1

Q4 สำรวจข้อมูลเกี่ยวกับระยะทางของแต่ละสายการบิน

```
df %>%
    group_by(carrier) %>%
    summarise(n = n(), # จำนวนไฟลท์บิน
              sum_distance = sum(distance), # ผลรวมระยะทางในการบิน
              mean_distance = round(mean(distance, rm.na = TRUE),2), # ระยะทางเฉลี่ย
              max_distance = max(distance), # ระยะทางที่บินไกลที่สุด
              min_distance = min(distance)) %>% # ระยะทางที่บินใกล้ที่สุด
    arrange(desc(n))
```

A tibble: 16 × 6

carrier	n	sum_distance	mean_distance	max_distance	min_distance
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
UA	57782	88482811	1531.32	4963	116
B6	54049	57815654	1069.69	2586	173
EV	51108	28766906	562.87	1389	80
DL	47658	58999610	1237.98	2586	94
AA	31947	42913762	1343.28	2586	187
MQ	25037	14280468	570.37	1147	184
US	19831	11121739	560.83	2153	94
9E	17294	9163911	529.89	1587	94
WN	12044	12007523	996.97	2133	169
VX	5116	12787097	2499.43	2586	2248
FL	3175	2110700	664.79	762	397
AS	709	1703018	2402.00	2402	2402
F9	681	1103220	1620.00	1620	1620
YV	544	204782	376.44	544	96
HA	342	1704186	4983.00	4983	4983
OO	29	14769	509.28	1008	229

Q5 สำรวจวันที่มีค่าเฉลี่ยเวลาที่ใช้ในการบินมากที่สุดจำนวน 5 วัน

```
df %>%
    group_by (year, month, day) %>%
    summarise (avg_air_time = round( mean( air_time), 2) ) %>%
    arrange (desc(avg_air_time)) %>%
    head(5)
```

A grouped_df: 5 × 4

year	month	day	avg_air_time
<dbl>	<dbl>	<dbl>	<dbl>
2013	2	8	161.76
2013	10	22	124.70
2013	11	19	124.10
2013	11	4	123.72
2013	11	28	122.24

```
`summarise()` has grouped output by 'year', 'month'. You can override using the
```

RPostgeSQL

```
# waiting for install T^T
library(RPostgreSQL)
```

```
ERROR: Error in library(RPostgreSQL): there is no package called ‘RPostgreSQL’

Warning message in install.packages("RPostgreSQL"):
“installation of package ‘RPostgreSQL’ had non-zero exit status”
Updating HTML index of packages in '.Library'

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

```
# connect to elephantsql
con <- dbConnect(
  PostgreSQL(), # what driver
  host = "###",
  dbname = "###",
  port = 5432,
  user = "###",
  password = "###"
)
```

```
# create sample data
sample_flights <- flights[1:5, 1:5]
sample_flights
```

A tibble: 5 × 5

year	month	day	dep_time	sched_dep_time
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2013	1	1	517	515
2013	1	1	533	529
2013	1	1	542	540
2013	1	1	544	545
2013	1	1	554	600

```
# write data to server
dbWriteTable(con, "sample_flights", sample_flights)
```

```
# query all data from database
dbGetQuery("SELECT * FROM sample_flights")
```

```
# disconnect to server
dbDisconnect(con)
```