Data transformation

Working on nycflights13 Library

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 4.2.2
```

```
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(knitr)
```

There are 5 data sets in `nycflights13` library

- **airlines**: airline names.
- **airports**: airport metadata
- **flights**: flights data
- **planes**: plane metadata.
- **weather**: hourly weather data

Which route is the most popular?

```
flights %>%
  mutate(route = paste0(origin, '-', dest)) %>%
  count(route) %>%
  arrange(desc(n)) %>%
  head(5) %>%
  kable()
```

route	n
JFK-LAX	11262
LGA-ATL	10263
LGA-ORD	8857
JFK-SFO	8204
LGA-CLT	6168

What are the popular destinations in each month?

```
flights %>%
  group_by(month, dest) %>%
  summarize(n = n()) %>%
  arrange(desc(n)) %>%
  mutate(rank = 1:n()) %>%
  filter(rank <= 5) %>%
  arrange(month) %>%
  select(-n) %>%
  pivot_wider(names_from = 'rank',
              values_from = 'dest') %>%
  kable()
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

month	1	2	3	4	5
1	ATL	ORD	BOS	MCO	FLL
2	ATL	ORD	BOS	MCO	FLL
3	ATL	ORD	BOS	MCO	FLL
4	ATL	ORD	LAX	BOS	MCO
5	ORD	ATL	LAX	BOS	SFO
6	ORD	ATL	LAX	BOS	SFO
7	ORD	ATL	LAX	BOS	CLT
8	ORD	ATL	LAX	BOS	SFO
9	ORD	LAX	ATL	BOS	CLT
10	ORD	ATL	LAX	BOS	CLT
11	ATL	ORD	LAX	BOS	CLT
12	ATL	LAX	MCO	ORD	CLT

Which airline had the most arrival delays?

Criteria: - Count only airlines with more than 10,000 flights - Arrival delays more than 10 mins are considered to be late

```
f <- flights %>% mutate(late = arr_delay > 10)
pct_fmt <- label_percent(accuracy = 0.01)

f %>%
  filter(!is.na(late)) %>%
  group_by(carrier, late) %>%
  summarize(n = n()) %>%
  mutate(pct = pct_fmt(n/sum(n))) %>%
  filter(sum(n) > 10000, late) %>%
  ungroup() %>%
  inner_join(airlines, by = 'carrier') %>%
  select(name, n, pct) %>%
  arrange(desc(pct)) %>%
  kable()
```

```
## `summarise()` has grouped output by 'carrier'. You can override using the
## `.groups` argument.
```

name	n	pct
ExpressJet Airlines Inc.	18202	35.61%
Envoy Air	8080	32.27%
JetBlue Airways	16528	30.58%
Southwest Airlines Co.	3553	29.50%
Endeavor Air Inc.	4879	28.21%
United Air Lines Inc.	15061	26.07%
American Airlines Inc.	7166	22.43%
US Airways Inc.	4449	22.43%
Delta Air Lines Inc.	10598	22.24%

Which month had the most arrival delays?

```
f %>%
  filter(!is.na(late)) %>%
  group_by(month, late) %>%
  summarize(n = n()) %>%
  mutate(pct = pct_fmt(n/sum(n))) %>%
  filter(late) %>%
  select(month, n, pct) %>%
  arrange(desc(pct)) %>%
  kable()
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

month	n	pct
12	10340	38.27%
7	10018	35.41%
6	9392	34.69%
4	8933	32.41%
8	8113	28.21%
2	6618	28.03%
1	7253	27.48%
3	7543	27.03%
5	7198	25.59%
11	5871	21.77%
10	6029	21.07%
9	4313	15.97%

How many flights were flying in each hour?

```
add_time_slots <- function(df) {
  for (i in 0:23) {
    col_name = sprintf('%02d-%02d', i, i+1)
    df[col_name] <- floor(df$dep_time / 100) > i &
                    ceiling(df$arr_time / 100) <= i+1
  }
  return(df)
}

f %>%
  add_time_slots() %>%
  select(21:ncol(.)) %>%
  colSums(na.rm = TRUE) %>%
  tibble(time_slot=names(.), count=.) %>%
  kable()
```

time_slot	count
00-01	6970
01-02	9203
02-03	9891
03-04	10344
04-05	10612
05-06	10629
06-07	10630
07-08	10631
08-09	10631
09-10	10631
10-11	10632
11-12	10633
12-13	10633
13-14	10633
14-15	10632
15-16	10631
16-17	10627

time_slot	count
17-18	10619
18-19	10587
19-20	10417
20-21	9168
21-22	5565
22-23	2608
23-24	29