problemset4

junyi zhang

GitHub

https://github.com/juny1z/Problemset4.git

Problem 1

a. Departure delay table

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

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
departure_delay <- nycflights13::flights %>%
  filter(!is.na(dep_delay)) %>%
  group_by(origin) %>%
  summarize(
    mean_dep_delay = mean(dep_delay),
    median_dep_delay = median(dep_delay)
```

```
  ) %>%
  left_join(airports, by = c("origin" = "faa")) %>%
  select(name, mean_dep_delay, median_dep_delay) %>%
  arrange(desc(mean_dep_delay)) %>%
  print(n = Inf)
```

```
# A tibble: 3 x 3
  name                mean_dep_delay median_dep_delay
  <chr>                        <dbl>            <dbl>
1 Newark Liberty Intl           15.1               -1
2 John F Kennedy Intl           12.1               -1
3 La Guardia                    10.3               -3
```

a. Arrival delay table

```
arrival_delay <- nycflights13::flights %>%
  filter(!is.na(arr_delay)) %>%
  group_by(dest) %>%
  filter(n() >= 10) %>%
  summarize(
    mean_arr_delay = mean(arr_delay),
    median_arr_delay = median(arr_delay)
  ) %>%
  left_join(airports, by = c("dest" = "faa")) %>%
  select(name, mean_arr_delay, median_arr_delay) %>%
  arrange(desc(mean_arr_delay)) %>%
  print(n = Inf)
```

```
# A tibble: 102 x 3
    name                                mean_arr_delay median_arr_delay
    <chr>                                        <dbl>            <dbl>
  1 "Columbia Metropolitan"                       41.8               28
  2 "Tulsa Intl"                                  33.7               14
  3 "Will Rogers World"                           30.6               16
  4 "Jackson Hole Airport"                        28.1               15
  5 "Mc Ghee Tyson"                               24.1                2
  6 "Dane Co Rgnl Truax Fld"                      20.2                1
  7 "Richmond Intl"                               20.1                1
  8 "Akron Canton Regional Airport"               19.7                3
  9 "Des Moines Intl"                             19.0                0
 10 "Gerald R Ford Intl"                          18.2                1
```

11	"Birmingham Intl"	16.9	-2
12	"Theodore Francis Green State"	16.2	1
13	"Greenville-Spartanburg International"	15.9	-0.5
14	"Cincinnati Northern Kentucky Intl"	15.4	-3
15	"Savannah Hilton Head Intl"	15.1	-1
16	"Manchester Regional Airport"	14.8	-3
17	"Eppley Afld"	14.7	-2
18	"Yeager"	14.7	-1.5
19	"Kansas City Intl"	14.5	0
20	"Albany Intl"	14.4	-4
21	"General Mitchell Intl"	14.2	0
22	"Piedmont Triad"	14.1	-2
23	"Washington Dulles Intl"	13.9	-3
24	"Cherry Capital Airport"	13.0	-10
25	"James M Cox Dayton Intl"	12.7	-3
26	"Louisville International Airport"	12.7	-2
27	"Chicago Midway Intl"	12.4	-1
28	"Sacramento Intl"	12.1	4
29	"Jacksonville Intl"	11.8	-2
30	"Nashville Intl"	11.8	-2
31	"Portland Intl Jetport"	11.7	-4
32	"Greater Rochester Intl"	11.6	-5
33	"Hartsfield Jackson Atlanta Intl"	11.3	-1
34	"Lambert St Louis Intl"	11.1	-3
35	"Norfolk Intl"	10.9	-4
36	"Baltimore Washington Intl"	10.7	-5
37	"Memphis Intl"	10.6	-2.5
38	"Port Columbus Intl"	10.6	-3
39	"Charleston Afb Intl"	10.6	-4
40	"Philadelphia Intl"	10.1	-3
41	"Raleigh Durham Intl"	10.1	-3
42	"Indianapolis Intl"	9.94	-3
43	"Charlottesville-Albemarle"	9.5	-5
44	"Cleveland Hopkins Intl"	9.18	-5
45	"Ronald Reagan Washington Natl"	9.07	-2
46	"Burlington Intl"	8.95	-4
47	"Buffalo Niagara Intl"	8.95	-5
48	"Syracuse Hancock Intl"	8.90	-5
49	"Denver Intl"	8.61	-2
50	"Palm Beach Intl"	8.56	-3
51	<NA>	8.25	-1
52	"Bob Hope"	8.18	-3
53	"Fort Lauderdale Hollywood Intl"	8.08	-3

54	"Bangor Intl"	8.03	-9
55	"Asheville Regional Airport"	8.00	-1
56	<NA>	7.87	0
57	"Pittsburgh Intl"	7.68	-5
58	"Gallatin Field"	7.6	-2
59	"NW Arkansas Regional"	7.47	-2
60	"Tampa Intl"	7.41	-4
61	"Charlotte Douglas Intl"	7.36	-3
62	"Minneapolis St Paul Intl"	7.27	-5
63	"William P Hobby"	7.18	-4
64	"Bradley Intl"	7.05	-10
65	"San Antonio Intl"	6.95	-9
66	"South Bend Rgnl"	6.5	-3.5
67	"Louis Armstrong New Orleans Intl"	6.49	-6
68	"Key West Intl"	6.35	7
69	"Eagle Co Rgnl"	6.30	-4
70	"Austin Bergstrom Intl"	6.02	-5
71	"Chicago Ohare Intl"	5.88	-8
72	"Orlando Intl"	5.45	-5
73	"Detroit Metro Wayne Co"	5.43	-7
74	"Portland Intl"	5.14	-5
75	"Nantucket Mem"	4.85	-3
76	"Wilmington Intl"	4.64	-7
77	"Myrtle Beach Intl"	4.60	-13
78	"Albuquerque International Sunport"	4.38	-5.5
79	"George Bush Intercontinental"	4.24	-5
80	"Norman Y Mineta San Jose Intl"	3.45	-7
81	"Southwest Florida Intl"	3.24	-5
82	"San Diego Intl"	3.14	-5
83	"Sarasota Bradenton Intl"	3.08	-5
84	"Metropolitan Oakland Intl"	3.08	-9
85	"General Edward Lawrence Logan Intl"	2.91	-9
86	"San Francisco Intl"	2.67	-8
87	<NA>	2.52	-6
88	"Yampa Valley"	2.14	2
89	"Phoenix Sky Harbor Intl"	2.10	-6
90	"Montrose Regional Airport"	1.79	-10.5
91	"Los Angeles Intl"	0.547	-7
92	"Dallas Fort Worth Intl"	0.322	-9
93	"Miami Intl"	0.299	-9
94	"Mc Carran Intl"	0.258	-8
95	"Salt Lake City Intl"	0.176	-8
96	"Long Beach"	-0.0620	-10

```
 97 "Martha\\\\'s Vineyard"                    -0.286            -11
 98 "Seattle Tacoma Intl"                      -1.10             -11
 99 "Honolulu Intl"                            -1.37              -7
100  <NA>                                     -3.84              -9
101 "John Wayne Arpt Orange Co"               -7.87             -11
102 "Palm Springs Intl"                      -12.7              -13.5
```

b. Aircraft model with the fastest average speed

```r
fastest_aircraft <- nycflights13::flights %>%
  filter(!is.na(air_time) & !is.na(distance)) %>%
  mutate(speed = distance / (air_time / 60)) %>%
  group_by(tailnum) %>%
  summarize(
    ave_speed = mean(speed),
    num_flights = n()
  ) %>%
  arrange(desc(ave_speed)) %>%
  slice(1) %>%
  left_join(planes, by = "tailnum") %>%
  select(model, ave_speed, num_flights)

print(fastest_aircraft)
```

```
# A tibble: 1 x 3
  model   ave_speed num_flights
  <chr>       <dbl>       <int>
1 777-222      501.           1
```

Problem 2

```r
# This block intentionally produces an error
stop("This is an intentional error.")
```

```
Error: This is an intentional error.
```

```r
get_temp <- function(month, year, data, celsius = FALSE, average_fn = mean){

  if (is.character(month)) {
    month <- match(tolower(month), tolower(c(month.abb, month.name)))
    if (is.na(month)) stop("Invalid month name")
  } else if (is.numeric(month) && (month < 1 || month > 12)) {
    stop("Invalid month")
  }

  if (!is.numeric(year))
    stop("Invalid year")

  temp_data <- data %>%
    filter(year == !!year, month_numeric == !!month)

  if (nrow(temp_data) == 0)
    stop("Invalid data")

  ave_temp <- temp_data %>%
    summarize(ave_temp = average_fn(temp, na.rm = TRUE)) %>%
    pull(ave_temp)

  if (celsius) {
    ave_temp <- (ave_temp - 32) * (5 / 9)
  }

  return(ave_temp)
}

nnmaps <- read.csv("/Users/zjyyy/Desktop/chicago-nmmaps.csv")

get_temp("Apr", 1999, data = nnmaps)
```

```
[1] 49.8
```

```r
get_temp("Apr", 1999, data = nnmaps, celsius = TRUE)
```

```
[1] 9.888889
```

```r
get_temp(10, 1998, data = nnmaps, average_fn = median)
```

```
[1] 55
```

```r
get_temp(13, 1998, data = nnmaps)
```

```
Error in get_temp(13, 1998, data = nnmaps): Invalid month
```

```r
get_temp(2, 2005, data = nnmaps)
```

```
Error in get_temp(2, 2005, data = nnmaps): Invalid data
```

```r
get_temp("November", 1999, data =nnmaps, celsius = TRUE,
         average_fn = function(x) {
           x %>% sort -> x
           x[2:(length(x) - 1)] %>% mean %>% return
         })
```

```
Error in get_temp("November", 1999, data = nnmaps, celsius = TRUE, average_fn = function(x) {
```

Problem 3

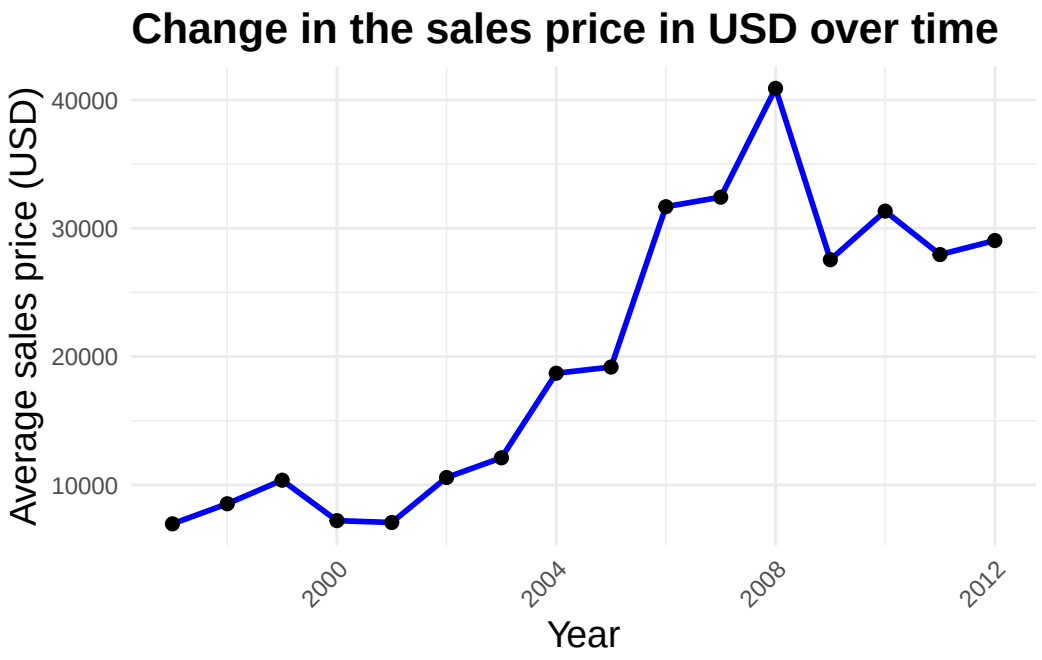
a. Is there a change in the sales price in USD over time?

```r
library(tidyverse)
library(ggplot2)
data <- read.csv("/Users/zjyyy/Desktop/df_for_ml_improved_new_market.csv")

data_ave_price <- data %>%
  group_by(year) %>%
  summarize(ave_price = mean(price_usd, na.rm = TRUE))

ggplot(data_ave_price, aes(x = year, y = ave_price)) +
  geom_line(color = "blue", linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Change in the sales price in USD over time",
    x = "Year",
```

```
    y = "Average sales price (USD)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



There is a clear change in the average sales price over time. There was a significant increase in sales prices until 2008, after which prices slightly leveled off or decreased.

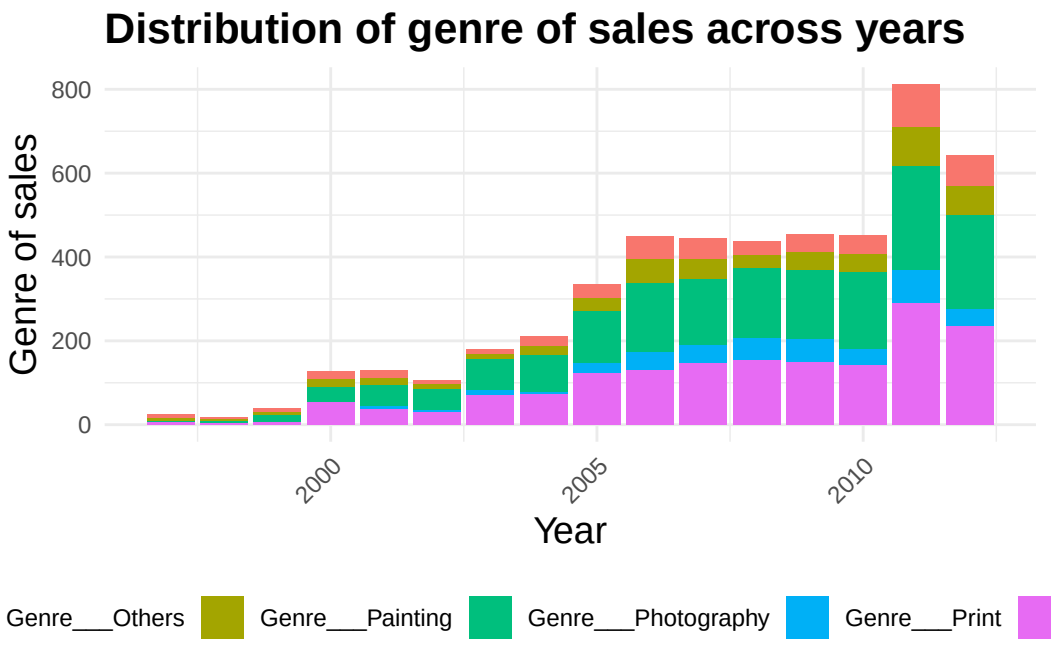
b. Does the distribution of genre of sales across years appear to change?

```
sales_genre <- c("Genre___Photography", "Genre___Print", "Genre___Sculpture", "Genre___Painti

data_sales_genre <- data %>%
  pivot_longer(cols = all_of(sales_genre),
               names_to = "genre",
               values_to = "sold") %>%
  filter(sold == 1)

ggplot(data_sales_genre, aes(x = year, fill = genre)) +
```

```
  geom_bar(position = "stack") +
  labs(
    title = "Distribution of genre of sales across years",
    x = "Year",
    y = "Genre of sales"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom"
  )
```



The distribution of genres changes significantly over time. In the earlier years, the volume of sales across genres is relatively low. However, from around 2000, the number of sales across all genres increases.

c. How does the genre affect the change in sales price over time?

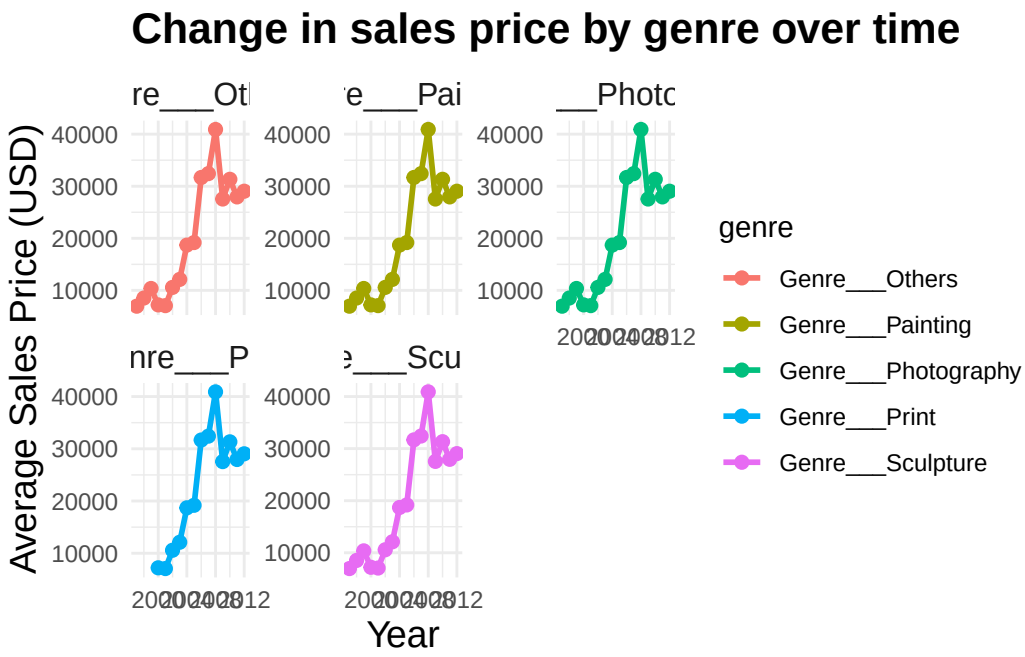
```
data_ave_price <- data %>%
  group_by(year) %>%
  summarize(ave_price = mean(price_usd, na.rm = TRUE))
```

```r
data_genre_price <- data_sales_genre %>%
  left_join(data_ave_price, by = "year") %>%
  group_by(year, genre) %>%
  summarize(Average_price = mean(ave_price, na.rm = TRUE))
```

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

```r
ggplot(data_genre_price, aes(x = year, y = Average_price, color = genre)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  facet_wrap(~ genre, scales = "free_y") +
  labs(
    title = "Change in sales price by genre over time",
    x = "Year",
    y = "Average Sales Price (USD)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    strip.text = element_text(size = 12)
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Each genre seems to follow a similar trend of rising prices until around 2008. However, certain genres like photography and painting appear to have achieved higher peaks compared to others.