Data Transformation

Jiruspak Franc

2024-07-24

```
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 become errors
```

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

1. Analysis of the most delayed flights

```
flights %>%
  arrange(desc(dep_delay)) %>%
  select(year, month, day, dep_time, sched_dep_time, dep_delay, carrier, flight, origin, dest) %>%
  head(10)
```

```
## # A tibble: 10 x 10
##     year month   day dep_time sched_dep_time dep_delay carrier flight origin
##    <int> <int> <int>    <int>          <int>     <dbl> <chr>    <int> <chr>
##  1  2013     1     9      641            900      1301 HA          51 JFK
##  2  2013     6    15     1432           1935      1137 MQ        3535 JFK
##  3  2013     1    10     1121           1635      1126 MQ        3695 EWR
##  4  2013     9    20     1139           1845      1014 AA         177 JFK
##  5  2013     7    22      845           1600      1005 MQ        3075 JFK
##  6  2013     4    10     1100           1900       960 DL        2391 JFK
##  7  2013     3    17     2321            810       911 DL        2119 LGA
##  8  2013     6    27      959           1900       899 DL        2007 JFK
##  9  2013     7    22     2257            759       898 DL        2047 LGA
## 10  2013    12     5      756           1700       896 AA         172 EWR
## # i 1 more variable: dest <chr>
```

Looking at the top 10 most delayed flights helps us identify which flights frequently encounter issues.

2. Analysis of the average flight time for each airline

```
flights %>%
  group_by(carrier) %>%
  summarize(avg_air_time = mean(air_time, na.rm = TRUE)) %>%
  arrange(desc(avg_air_time)) %>%
  left_join(airlines, by = "carrier")
```

```
## # A tibble: 16 x 3
##    carrier avg_air_time name
##    <chr>          <dbl> <chr>
##  1 HA             623.  Hawaiian Airlines Inc.
##  2 VX             337.  Virgin America
##  3 AS             326.  Alaska Airlines Inc.
##  4 F9             230.  Frontier Airlines Inc.
##  5 UA             212.  United Air Lines Inc.
##  6 AA             189.  American Airlines Inc.
##  7 DL             174.  Delta Air Lines Inc.
##  8 B6             151.  JetBlue Airways
##  9 WN             148.  Southwest Airlines Co.
## 10 FL             101.  AirTran Airways Corporation
## 11 MQ              91.2 Envoy Air
## 12 EV              90.1 ExpressJet Airlines Inc.
## 13 US              88.6 US Airways Inc.
## 14 9E              86.8 Endeavor Air Inc.
## 15 OO              83.5 SkyWest Airlines Inc.
## 16 YV              65.7 Mesa Airlines Inc.
```

Calculating the average flight time for each airline helps to see the differences in average flight times between airlines.

3. Analysis of how arrival delays are related to weather conditions

```
combined_data <- flights %>%
  left_join(weather, by = c("year", "month", "day", "hour", "origin")) %>%
  select(year, month, day, carrier, arr_delay, temp, dewp, humid, wind_speed, precip)

combined_data %>%
  group_by(year, month, carrier) %>%
  summarise(
    avg_arr_delay = mean(arr_delay, na.rm = TRUE),
    avg_temp = mean(temp, na.rm = TRUE),
    avg_dewp = mean(dewp, na.rm = TRUE),
    avg_humid = mean(humid, na.rm = TRUE),
    avg_wind_speed = mean(wind_speed, na.rm = TRUE),
    avg_precip = mean(precip, na.rm = TRUE)
  )
```

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

```
## # A tibble: 185 x 9
## # Groups:   year, month [12]
##     year month carrier avg_arr_delay avg_temp avg_dewp avg_humid avg_wind_speed
##    <int> <int> <chr>           <dbl>    <dbl>    <dbl>     <dbl>          <dbl>
##  1  2013     1 9E             10.2       36.5     22.0     58.7          12.1
##  2  2013     1 AA              0.982     36.4     22.0     58.5          12.0
##  3  2013     1 AS              8.97      34.8     22.7     63.9           8.98
##  4  2013     1 B6              4.72      36.2     22.2     59.5          12.1
##  5  2013     1 DL             -4.40      36.5     21.9     58.0          12.1
##  6  2013     1 EV             25.2       36.8     22.3     58.6          10.5
##  7  2013     1 F9             21.8       35.9     22.0     59.5          11.9
##  8  2013     1 FL              3.32      36.5     21.8     57.5          11.8
##  9  2013     1 HA             27.5       35.8     23.4     62.3          12.5
## 10  2013     1 MQ              7.88      36.9     21.8     56.8          12.0
## # i 175 more rows
## # i 1 more variable: avg_precip <dbl>
```

It helps understand how weather conditions in each month affect the delays of flights for each airline.

4. Examining the aircraft models and seat counts for each airline

```
flight_planes <- flights %>%
  left_join(planes, by = "tailnum") %>%
  select(carrier, tailnum, type, model, seats)

flight_planes_airlines <- flight_planes %>%
  left_join(airlines, by = "carrier")

flight_planes_airlines %>%
  group_by(carrier, name, type, model) %>%
  summarise(
    avg_seats = mean(seats, na.rm = TRUE)
  ) %>%
  arrange(carrier, model)
```

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

```
## # A tibble: 155 x 5
## # Groups:   carrier, name, type [32]
##    carrier name                   type                     model       avg_seats
##    <chr>   <chr>                  <chr>                    <chr>           <dbl>
##  1 9E      Endeavor Air Inc.      Fixed wing multi engine  CL-600-2B19        55
##  2 9E      Endeavor Air Inc.      Fixed wing multi engine  CL-600-2D24        95
##  3 9E      Endeavor Air Inc.      <NA>                     <NA>              NaN
##  4 AA      American Airlines Inc. Fixed wing single engine 150                 2
##  5 AA      American Airlines Inc. Fixed wing single engine 172E                4
##  6 AA      American Airlines Inc. Fixed wing single engine 172M                4
##  7 AA      American Airlines Inc. Rotorcraft               206B                5
##  8 AA      American Airlines Inc. Fixed wing single engine 210-5(205)          6
##  9 AA      American Airlines Inc. Rotorcraft               230                11
## 10 AA      American Airlines Inc. Fixed wing multi engine  310Q                6
## # i 145 more rows
```

If certain models have a higher or significantly higher number of seats in some airlines but are lower or not present in others, it may reflect different aircraft selection based on the airline's needs or strategy.

5. Analysis of which airlines have aircraft models and types with the longest flight distances

```r
flight_planes <- flights %>%
  left_join(planes, by = "tailnum") %>%
  select(carrier, tailnum, model, type, distance)

flight_planes_airlines <- flight_planes %>%
  left_join(airlines, by = "carrier")

flight_planes_airlines %>%
  group_by(carrier, name, model, type) %>%
  summarise(
    avg_distance = mean(distance, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  arrange(desc(avg_distance))
```

```
## # A tibble: 155 x 5
##    carrier name                   model     type                    avg_distance
##    <chr>   <chr>                  <chr>     <chr>                          <dbl>
##  1 HA      Hawaiian Airlines Inc. A330-243  Fixed wing multi engine        4983
##  2 UA      United Air Lines Inc.  767-424ER Fixed wing multi engine        3850.
##  3 VX      Virgin America         A319-115  Fixed wing multi engine        2525.
##  4 VX      Virgin America         A319-112  Fixed wing multi engine        2523.
##  5 UA      United Air Lines Inc.  777-222   Fixed wing multi engine        2520
##  6 VX      Virgin America         A320-214  Fixed wing multi engine        2498.
##  7 DL      Delta Air Lines Inc.   757-212   Fixed wing multi engine        2475
##  8 DL      Delta Air Lines Inc.   A330-323  Fixed wing multi engine        2422
##  9 AS      Alaska Airlines Inc.   737-890   Fixed wing multi engine        2402
## 10 AS      Alaska Airlines Inc.   737-8FH   Fixed wing multi engine        2402
## # i 145 more rows
```

- If an airline has aircraft models or types with the longest flight distances, it indicates that the airline may focus on providing long-haul flights or has aircraft designed for long-distance travel.
- This analysis also helps to see that some airlines may choose aircraft with long-range capabilities to cover longer routes or meet specific service requirements.