cluster-destination-timeofday

```
library(conflicted)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
```

```
library(rlang)
library(gt)
  library(gtExtras)
conflicts_prefer(
  dplyr::filter
)
```

```
## [conflicted] Will prefer dplyr::filter over any other package.
```

```
unzip("./passenger_data.zip")
```

```
data = "./passenger_data.csv" |>
    read_csv(
        col_types = cols(
            X = col_integer(),
            Airfield = col_factor(),
            S2 = col_datetime(),
            Wait_Time = col_integer(),
            C_Start = col_integer(),
            C0 = col_integer(),
            C_avg = col_double(),
            Sch_Departure = col_datetime(),
            Act_Departure = col_datetime(),
            BFO_Dest_City = col_factor(),
            BFO_Destination_Country_Code = col_factor(),
            order = col_integer(),
            Pass_ID = col_integer(),
            Departure_Date = col_date(),
            Departure_Time = col_time(),
            Time_of_Day = col_factor(),
            Period_of_Week = col_factor(),
            Day_of_Week = col_factor(),
            Month = col_factor(),
            Season = col_factor(),
```

```
            Year = col_factor(),
            Flight_ID = col_integer(),
            Delay_in_Seconds = col_integer(),
            .default = col_guess()
        )
    )
```

```
# ~1500 observations from Airfield SAF make it unsuitable for clustering
# the destinations below have too few observations
data = data |>
  filter(
    Airfield == "AUC",
    BFO_Destination_Country_Code != "WIC",
    BFO_Dest_City != "BOR008",
    Year == "2028" # otherwise only 4 obs for 2030
  ) |>
  filter(S2 > ymd_hms("2028-08-31 00:00:00")) |>
  mutate(
    BFO_Dest_City_or_CC = if_else(
      BFO_Destination_Country_Code == "BOR",
      true = paste(
        "City",
        BFO_Dest_City,
        sep = "-"
      ),
      false = paste(
        "Country",
        BFO_Destination_Country_Code,
        sep = "-"
      )
    ) |>
      as_factor(),
    .after = BFO_Destination_Country_Code
  ) |>
  unite(
    col = cluster,
    Time_of_Day,
    BFO_Dest_City_or_CC,
    sep = "_",
    remove = FALSE,
    na.rm = FALSE
  ) |>
  mutate(cluster = as_factor(cluster))
```

```
summary(data)
```

```
##  original_frame        Pass_ID              X           Airfield
##  Length:191363      Min.   :5345525   Min.   : 104917   AUC:191363
##  Class :character   1st Qu.:5658913   1st Qu.: 288980   SAF:     0
##  Mode  :character   Median :5939159   Median : 646178
##                     Mean   :5924711   Mean   : 610061
##                     3rd Qu.:6199234   3rd Qu.: 869486
##                     Max.   :6438752   Max.   :1098324
```

```
##                    NA's   :5         NA's   :77620
##        S2                            Wait_Time        C_Start
##  Min.   :2028-08-31 18:54:00.00   Min.   : 1.00   Min.   :0.000
##  1st Qu.:2028-10-04 08:02:00.00   1st Qu.: 2.00   1st Qu.:1.000
##  Median :2028-11-03 13:13:00.00   Median : 4.00   Median :1.000
##  Mean   :2028-11-03 05:03:21.86   Mean   : 5.05   Mean   :1.302
##  3rd Qu.:2028-12-04 12:50:00.00   3rd Qu.: 6.00   3rd Qu.:2.000
##  Max.   :2028-12-31 20:26:00.00   Max.   :75.00   Max.   :3.000
##                                   NA's   :28441   NA's   :16717
##        C0            C_avg       Sch_Departure
##  Min.   :1.000   Min.   :0.500   Min.   :2028-08-31 20:16:00.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2028-10-04 09:46:00.00
##  Median :1.000   Median :1.000   Median :2028-11-03 14:51:00.00
##  Mean   :1.338   Mean   :1.305   Mean   :2028-11-03 06:47:03.00
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2028-12-04 14:46:00.00
##  Max.   :3.000   Max.   :3.000   Max.   :2028-12-31 20:16:00.00
##                  NA's   :16717   NA's   :114
##  Act_Departure                     BFO_Dest_City   BFO_Destination_Country_Code
##  Min.   :2028-08-31 20:16:00.00   QUE    :87207   BOR:136953
##  1st Qu.:2028-10-04 09:16:00.00   CWL    :40188   VES: 39041
##  Median :2028-11-03 14:48:00.00   VES033 :24800   NEN:  4418
##  Mean   :2028-11-03 06:54:49.97   VES064 :14123   SCO: 10951
##  3rd Qu.:2028-12-04 14:51:00.00   SCO032 :10498   WIC:     0
##  Max.   :2028-12-31 21:50:00.00   SAF    : 9558
##                                   (Other): 4989
##           cluster       BFO_Dest_City_or_CC     order
##  3_City-QUE   :39905   City-CWL   :40188    Min.   :5372562
##  4_City-QUE   :30105   City-QUE   :87207    1st Qu.:5685624
##  3_Country-VES:21102   Country-VES:39041    Median :5960040
##  3_City-CWL   :18211   Country-NEN: 4418    Mean   :5945902
##  2_City-QUE   :17043   Country-SCO:10951    3rd Qu.:6214482
##  2_City-CWL   :12710   City-SAF   : 9558    Max.   :6459925
##  (Other)      :52287
##  Departure_Date       Departure_Time    Time_of_Day     Period_of_Week
##  Min.   :2028-08-31   Length:191363     2:40315     1 - WEEKDAY:134138
##  1st Qu.:2028-10-04   Class1:hms        3:81485     2 - WEEKEND: 57225
##  Median :2028-11-03   Class2:difftime   4:69407
##  Mean   :2028-11-02   Mode  :numeric    1:  156
##  3rd Qu.:2028-12-04
##  Max.   :2028-12-31
##
##   Day_of_Week    Month      Season       Year          Flight_ID
##  5 - FRI:24246   9 :43114   3: 43125   2028:191363   Min.   :18102
##  6 - SAT:31123   10:48752   4:148238   2030:     0   1st Qu.:19118
##  7 - SUN:26102   11:46575   1:     0                 Median :20043
##  1 - MON:21513   12:52911                           Mean   :20005
##  2 - TUE:24548   8 :   11                           3rd Qu.:20922
##  3 - WED:30581   1 :    0                           Max.   :21678
##  4 - THU:33250                                      NA's   :113743
##  Delay_in_Seconds
##  Min.   :-1440
##  1st Qu.:    0
##  Median :    0
##  Mean   : 1097
```

```
##  3rd Qu.: 1080
##  Max.   :52980
##  NA's   :113743
```

```
cl = data |>
  group_by(cluster) |>
  summarise(n = n())
cl |>
  mutate(
    count = n,
    rate = n/732, # number of hours: 732 apiece
    .keep = "unused",
    .after = cluster
  ) |>
  arrange(cluster) |>
  gt() |>
  tab_header(
    title = "Passenger arrival rates at **AUC**" |>
      md(),
    subtitle = "clustered by destination and time of day"
  )
```

Passenger arrival rates at **AUC**

clustered by destination and time of day

cluster	count	rate
2_City-CWL	12710	17.36338798
3_City-QUE	39905	54.51502732
3_Country-VES	21102	28.82786885
4_City-QUE	30105	41.12704918
3_City-CWL	18211	24.87841530
4_Country-NEN	3354	4.58196721
4_Country-VES	11231	15.34289617
4_City-CWL	9267	12.65983607
3_Country-SCO	1267	1.73087432
1_City-QUE	154	0.21038251
4_Country-SCO	9354	12.77868852
2_Country-VES	6706	9.16120219
2_City-QUE	17043	23.28278689
4_City-SAF	6096	8.32786885
2_City-SAF	2471	3.37568306
2_Country-SCO	330	0.45081967
3_City-SAF	991	1.35382514
1_Country-VES	2	0.00273224
2_Country-NEN	1055	1.44125683
3_Country-NEN	9	0.01229508