French Train stats

We've focused a lot on US data over the past few weeks, so many thanks to Mathilda for helping curate this week's French trains dataset! She posted an awesome visualization of this data - if you are interested in a heat-map style `geom_tile()`, take a look at her post https://twitter.com/noccaea/status/1095735292206739456.

The data comes from the SNCF open data portal - there are additional datasets there you can download, but fair to say most things are in French! :smile:

The SNCF (National Society of French Railways) is France's national state-owned railway company. Founded in 1938, it operates the country's national rail traffic along with Monaco, including the TGV, France's high-speed rail network. This dataset covers 2015-2018 with a lot of different train stations. The dataset primarily covers aggregate trip times, delay times, cause for delay, etc for each station - lots of different ways to approach the **full_trains.csv** dataset with it's 27 columns!

If you are interested in a shorter data journey, check out the small_trains.csv dataset - it is only 13 columns and has a `gather()` already performed.

Lastly, if for some reason you'd like to see the raw untranslated dataset it is also available here.

Grab the raw data here

```
trains_raw <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data
small_trains <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

Data Dictionary

small_trains.csv

variable	class	description
year	integer/date	Year of Observation
month	double/date	Month of Observation
service	factor	Type of train service (National, Internation, NA)
departure_station	character	Departure Station (name)
arrival_station	character	Arrival Station (name)
journey_time_avg	double	Average Journey time (minutes)
total_num_trips	double	Total number of trains in the time period
avg_delay_all_departing	double	The average delay (minutes) for all departing trains
avg_delay_all_arriving	double	Average delay (minutes) for all arriving trains
num_late_at_departure	double	Number of trains that were late at departure
num_arriving_late	double	Number of trains arriving late
delay_cause	character	Cause for delay
delayed_number	double	Percent of trains delayed

full_trains.csv

variable	class	description
year	integer/date	Year of Observation
month	double/date	Month of Observation
service	factor	Type of train service (National, Internation, NA)
departure_station	character	Departure Station (name)
arrival_station	character	Arrival Station (name)
journey_time_avg	double	Average Journey time (minutes)
total_num_trips	double	Total number of trains in the time period
num_of_canceled_trains	double	Number of canceled trains
comment_cancellations	character	Comment for Cancellations
num_late_at_departure	double	Number of trains that were late at departure
avg_delay_late_at_departure	double	The average delay (minutes) for trains late at departure
avg_delay_all_departing	double	The average delay (minutes) for all departing trains
comment_delays_at_departure	character	Comment for trains delayed at departure
num_arriving_late	double	Number of trains arriving late
avg_delay_late_on_arrival	double	Average delay (minutes) for trains that were late on arrival
avg_delay_all_arriving	double	Average delay (minutes) for all arriving trains
comment_delays_on_arrival	character	Comment for delayes on arrival
delay_cause_external_cause	double	Cause for delay (%) - External Cause
delay_cause_rail_infrastructure	double	Cause for delay (%) - Rail infrastructure
delay_cause_traffic_management	double	Cause for delay (%) - Traffic management
delay_cause_rolling_stock	double	Cause for delay (%) - Rolling stock
delay_cause_station_management	double	Cause for delay (%) - Station management

variable	class	description
delay_cause_travelers	double	Cause for delay (%) - Travelers
num_greater_15_min_late	double	Number of trains greater than 15 min late
avg_delay_late_greater_15_min	double	Average delay of trains that were late more than 15 min
num_greater_30_min_late	double	Number of trains greater than 30 min late
num_greater_60_min_late	double	Number of trains greater than 60 min late

Spoilers - Cleaning Script

```
library(tidyverse)

# translated col names
english_names <- c(
  "year", "month", "service", "departure_station", "arrival_station", "journey_time_avg",
  "total_num_trips", "num_of_canceled_trains", "comment_cancellations",
  "num_late_at_departure", "avg_delay_late_at_departure",
  "avg_delay_all_departing", "comment_delays_at_departure", "num_arriving_late",
  "avg_delay_late_on_arrival", "avg_delay_all_arriving",
  "comment_delays_on_arrival", "delay_cause_external_cause", "delay_cause_rail_infrastructure",
  "delay_cause_traffic_management", "delay_cause_rolling_stock",
  "delay_cause_station_management", "delay_cause_travelers", "num_greater_15_min_late",
  "avg_delay_late_greater_15_min", "num_greater_30_min_late",
  "num_greater_60_min_late", "period", "delay_for_external_cause",
  "delay_for_railway_infrastructure", "delay_for_traffic_managment", "delay_for_rolling_stock",
  "delay_for_station_management", "delay_for_passengers"
)

# read in the dataset
df <- read_delim(here::here("2019", "2019-02-19", "regularite-mensuelle-tgv-aqst.csv"),
                 delim = ";") %>%
  set_names(nm = english_names)

# select columns of interest and create dictionary
trains_df <- df %>%
  select(year:num_greater_60_min_late)

tomtom::create_dictionary(trains_df)

# write to csv
trains_df %>%
  write_csv("full_trains.csv")

# gather cause for delay and create a more focused dataset
small_df <- df %>%
```

```
  gather(delay_cause, delayed_number, delay_cause_external_cause:delay_cause_travelers) %>%
  select(year:total_num_trips, avg_delay_all_departing, avg_delay_all_arriving, num_late_at_departure, r

# create data dictionary
small_df %>%
  tomtom::create_dictionary()

# write to csv
small_df %>%
  write_csv("small_trains.csv")
```