Make Datasets

```r
library(tidyverse)
library(lubridate)
library(anytime)
library(matlab)
library(readr)
library(stringr)
```

Define where to get the data from and where to put the resulting data sets:

```r
data_source_dir = "../data_raw/"
data_destination_dir = "../datasets/"
```

Define helper functions that get information from a filename:

```r
get_provider_by_filename = function(filename) {
  split_result = str_split(filename, "_")
  return(split_result[[1]][2])
}

# example:
get_provider_by_filename("1544519617_vodafone_ul.txt")
```

```
## [1] "vodafone"
```

```r
get_datatype_by_filename = function(filename) {
  split_result = str_split(filename, "_")
  ending = split_result[[1]][3]
  split_result_ending = str_split(ending, "\\.")
  return(split_result_ending[[1]][1])
}

# example:
get_datatype_by_filename("1544519617_vodafone_ul.txt")
```

```
## [1] "ul"
```

```r
get_start_time_by_filename = function(filename) {
  split_result = str_split(filename, "_")
  result = as_datetime(as.integer(split_result[[1]][[1]]))
  return(result)
}

# example:
get_start_time_by_filename("1544519617_vodafone_ul.txt")
```

```
## [1] "2018-12-11 09:13:37 UTC"
```

This function creates a dataset for a given `datatype` like “ul”, “dl”, “context” or cells. The provider is added as an extra column as well as the scenario.

```r
make_dataset = function(data_location, datatype) {

  # read all datasets and store them in a list to combine them later
  datasets = list()

  scenarios = c("urban", "suburban", "campus", "highway")
  for (cur_scenario in scenarios) {

    # get the current path where the files are located and list the files
    cur_path = str_c(data_location, "/", cur_scenario)
    data_files = list.files(cur_path)

    # now read each file
    for (cur_filename in data_files) {

      # only read when the datatype matches the one we want
      cur_datatype = get_datatype_by_filename(cur_filename)
      if (cur_datatype != datatype) {
        next
      }

      # read the file and add a column for the provider and for the scenario
      cur_provider = get_provider_by_filename(cur_filename)
      cur_start_time = get_start_time_by_filename(cur_filename)
      cur_dataset = read_csv(str_c(cur_path, "/", cur_filename), col_type=cols()) %>%
        mutate(scenario=cur_scenario, provider=cur_provider, start_time=cur_start_time)

      # attach it to the list
      datasets[[length(datasets)+1]] = cur_dataset
    }
  }

  # build the final dataset and convert the seconds to proper dates
  final_dataset = bind_rows(datasets) %>%
    mutate(timestamp=as_datetime(timestamp_ms)) %>%
    arrange(start_time)
  return(final_dataset)
}
```

Now create the data sets:

```r
dataset_ul = make_dataset(data_source_dir, datatype="ul")
glimpse(dataset_ul)
```

```
## Rows: 6,180
## Columns: 28
## $ time_s            <dbl> 10.49, 21.71, 30.83, 42.03, 52.86, 65.00, 71.84, ...
## $ timestamp_ms      <dbl> 1544432937, 1544432948, 1544432957, 1544432968, 1...
## $ distance_m        <dbl> 100.93, 233.54, 316.60, 413.04, 515.22, 674.17, 7...
## $ latitude          <dbl> 51.49052, 51.49070, 51.49058, 51.49131, 51.49218,...
## $ longitude         <dbl> 7.413815, 7.415705, 7.416909, 7.417024, 7.416533,...
## $ altitude          <dbl> 161.41, 156.16, 156.66, 156.02, 155.88, 152.90, 1...
## $ velocity_mps      <dbl> 11.80, 11.49, 7.93, 10.44, 10.92, 12.02, 10.28, 0...
## $ acceleration_mpss <dbl> 0.13, -0.26, 0.23, 0.06, 0.56, 0.09, -1.25, 0.00,...
## $ direction         <dbl> 79.23, 85.64, 102.61, 349.23, 333.63, 348.79, 344...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -99, -97, -96, -82, -101, -106, -112, -99, -98, -...
## $ rsrq_db           <dbl> -9, -12, -12, -11, -14, -13, -18, -15, -15, -14, ...
## $ rssnr_db          <dbl> -1, -2, 5, 11, -3, -3, -6, -4, -6, -4, -6, -3, -2...
## $ cqi               <dbl> 8, 9, 5, 15, 6, 6, 3, 4, 7, 4, 4, 5, 6, 5, 1, 4, ...
## $ ss                <dbl> 36, 42, 42, 53, 39, 33, 31, 41, 40, 44, 43, 42, 4...
## $ ta                <dbl> 9, 7, 7, 7, 7, 7, 7, 12, 13, 13, 13, 13, 11, 13, ...
## $ ci                <dbl> 13828122, 13416987, 13416987, 13416987, 13416987,...
## $ pci               <dbl> 452, 62, 62, 62, 62, 62, 62, 34, 38, 38, 38, 37, ...
## $ id                <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,...
## $ payload_mb        <dbl> 1.0, 6.0, 5.0, 7.0, 5.0, 8.0, 9.0, 7.0, 10.0, 2.0...
## $ throughput_mbits  <dbl> 4.66, 3.97, 6.52, 1.37, 0.80, 1.04, 2.34, 4.09, 2...
## $ rtt_ms            <dbl> 47, 1493, 146, 1348, 5621, 679, 152, 444, 1387, 1...
## $ txPower_dbm       <dbl> 20.56, 21.12, 21.38, 18.78, 18.64, 21.14, 19.69, ...
## $ f_mhz             <dbl> 1750, 1750, 1750, 1750, 1750, 1750, 1750, 880, 88...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "o2", "o2", "o2", "o2", "o2", "o2", "...
## $ start_time        <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
## $ timestamp         <dttm> 2018-12-10 09:08:57, 2018-12-10 09:09:08, 2018-1...
```

```r
dataset_dl = make_dataset(data_source_dir, datatype="dl")
glimpse(dataset_dl)
```

```
## Rows: 6,516
## Columns: 27
## $ time_s            <dbl> 10.38, 28.46, 30.86, 47.81, 56.27, 63.97, 70.96, ...
## $ timestamp_ms      <dbl> 1544432936, 1544432955, 1544432957, 1544432974, 1...
## $ distance_m        <dbl> 100.93, 300.95, 316.60, 464.65, 562.62, 650.24, 7...
## $ latitude          <dbl> 51.49052, 51.49063, 51.49058, 51.49175, 51.49257,...
## $ longitude         <dbl> 7.413815, 7.416688, 7.416909, 7.416834, 7.416226,...
## $ altitude          <dbl> 161.41, 155.68, 156.66, 154.60, 155.12, 153.76, 1...
## $ velocity_mps      <dbl> 11.80, 8.02, 7.93, 10.08, 12.44, 12.03, 11.52, 0....
## $ acceleration_mpss <dbl> 0.13, 0.15, 0.23, -0.15, -0.16, -0.47, -0.25, 0.0...
## $ direction         <dbl> 79.23, 100.75, 102.61, 344.61, 337.81, 348.88, 34...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -99, -96, -96, -100, -101, -105, -112, -99, -98, ...
## $ rsrq_db           <dbl> -9, -12, -12, -13, -14, -15, -18, -15, -13, -14, ...
## $ rssnr_db          <dbl> -1, 5, 5, -1, -3, -4, -6, -4, -6, -4, -6, -3, -2,...
## $ cqi               <dbl> 8, 5, 5, 7, 5, 5, 3, 4, 6, 4, 5, 5, 6, 5, 1, 4, 6...
## $ ss                <dbl> 36, 42, 42, 38, 39, 36, 31, 41, 41, 44, 44, 42, 4...
## $ ta                <dbl> 9, 7, 7, 7, 7, 7, 7, 12, 13, 13, 13, 13, 11, 13, ...
## $ ci                <dbl> 13828122, 13416987, 13416987, 13416987, 13416987,...
## $ pci               <dbl> 452, 62, 62, 62, 62, 62, 62, 34, 38, 38, 38, 37, ...
## $ id                <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,...
## $ payload_mb        <dbl> 6.0, 10.0, 7.0, 7.0, 9.0, 3.0, 3.0, 0.5, 5.0, 2.0...
## $ throughput_mbits  <dbl> 2.38, 0.90, 1.09, 0.45, 0.51, 0.42, 0.71, 0.63, 0...
## $ rtt_ms            <dbl> 54, 1573, 144, 1346, 5662, 786, 151, 444, 1329, 1...
## $ f_mhz             <dbl> 1845, 1845, 1845, 1845, 1845, 1845, 1845, 850, 85...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "o2", "o2", "o2", "o2", "o2", "o2", "...
## $ start_time        <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
## $ timestamp         <dttm> 2018-12-10 09:08:56, 2018-12-10 09:09:15, 2018-1...
```

```r
dataset_context = make_dataset(data_source_dir, datatype="context")
glimpse(dataset_context)
```

```
## Rows: 71,027
## Columns: 22
## $ time_s            <dbl> 0.06, 1.07, 2.07, 3.07, 4.07, 5.07, 6.07, 7.07, 8...
## $ timestamp_ms      <dbl> 1544432926, 1544432927, 1544432928, 1544432929, 1...
## $ distance_m        <dbl> 0.00, 6.79, 14.43, 23.00, 33.08, 43.81, 54.75, 66...
## $ latitude          <dbl> 51.49033, 51.49036, 51.49038, 51.49040, 51.49042,...
## $ longitude         <dbl> 7.412292, 7.412422, 7.412551, 7.412716, 7.412865,...
## $ altitude          <dbl> 161.88, 162.54, 162.67, 162.96, 162.87, 162.82, 1...
## $ velocity_mps      <dbl> 6.76, 7.65, 8.57, 10.08, 10.73, 10.93, 11.19, 11....
## $ acceleration_mpss <dbl> 0.00, 0.89, 0.92, 1.51, 0.65, 0.19, 0.26, 0.47, 0...
## $ direction         <dbl> 76.84, 77.47, 76.36, 77.02, 77.88, 78.33, 79.23, ...
## $ isRegistered      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ rsrp_dbm          <dbl> -98, -101, -101, -94, -94, -98, -98, -94, -94, -9...
## $ rsrq_db           <dbl> -10, -12, -12, -9, -9, -8, -8, -9, -9, -9, -9, -9...
## $ rssnr_db          <dbl> -1, -1, -1, 5, 5, 1, 1, -2, -2, -2, -1, -1, -3, -...
## $ cqi               <dbl> 9, 6, 6, 12, 12, 10, 10, 5, 5, 5, 8, 8, 6, 6, 5, ...
## $ ss                <dbl> 37, 36, 36, 40, 40, 36, 36, 41, 41, 41, 36, 36, 3...
## $ ta                <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9...
## $ ci                <dbl> 13828122, 13828122, 13828122, 13828122, 13828122,...
## $ pci               <dbl> 452, 452, 452, 452, 452, 452, 452, 452, 452, 452,...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "o2", "o2", "o2", "o2", "o2", "o2", "...
## $ start_time        <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
## $ timestamp         <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:47, 2018-1...
```

```r
dataset_cells = make_dataset(data_source_dir, datatype="cells")
glimpse(dataset_cells)
```

```
## Rows: 93,443
## Columns: 22
## $ time_s            <dbl> 0.07, 0.07, 0.07, 1.07, 1.07, 2.07, 2.07, 3.07, 4...
## $ timestamp_ms      <dbl> 1544432926, 1544432926, 1544432926, 1544432927, 1...
## $ distance_m        <dbl> 0.02, 0.02, 0.02, 6.80, 6.81, 14.45, 14.46, 23.01...
## $ latitude          <dbl> 51.49033, 51.49033, 51.49033, 51.49036, 51.49036,...
## $ longitude         <dbl> 7.412292, 7.412292, 7.412292, 7.412422, 7.412422,...
## $ altitude          <dbl> 161.88, 161.88, 161.88, 162.54, 162.54, 162.67, 1...
## $ velocity_mps      <dbl> 6.76, 6.76, 6.76, 7.65, 7.65, 8.57, 8.57, 10.08, ...
## $ acceleration_mpss <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ direction         <dbl> 76.84, 76.84, 76.84, 77.47, 77.47, 76.36, 76.36, ...
## $ isRegistered      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ rsrp_dbm          <dbl> -99, -103, -103, -104, -107, -104, -107, -100, -1...
## $ rsrq_db           <dbl> -12, -14, -16, -14, -17, -14, -17, -17, -17, -11,...
## $ rssnr_db          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ cqi               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ss                <dbl> 39, 38, 39, 36, 35, 36, 35, 42, 42, 38, 38, 38, 3...
## $ ta                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ ci                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ pci               <dbl> 146, 450, 266, 450, 146, 450, 146, 450, 450, 450,...
## $ scenario          <chr> "campus", "campus", "campus", "campus", "campus",...
## $ provider          <chr> "o2", "o2", "o2", "o2", "o2", "o2", "o2", "o2", "...
## $ start_time        <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
## $ timestamp         <dttm> 2018-12-10 09:08:46, 2018-12-10 09:08:46, 2018-1...
```

Add the eNodeB-ID

This function converts a cell-ID to an eNodeB-ID:

```
cell_id_to_enodeb = function(cell_id) {
  result = tryCatch(
    {
      hex_string = as.character(as.hexmode(cell_id))
      enodeb_hex = str_sub(hex_string, start=1, end=-3)
      enodeb_integer = as.integer(as.hexmode(enodeb_hex))
      return(enodeb_integer)
    },
    error = function(err) {
      return(NA)
    }
  )
  return(result)
}
```

Let's test it using the example from the slides (the correct eNodeB-ID is 50464):

```
cell_id_to_enodeb(12918809)
```

```
## [1] 50464
```

A few more tests:

```
print(cell_id_to_enodeb(13828122)==54016)
```

```
## [1] TRUE
```

```
print(cell_id_to_enodeb(26385408)==103068)
```

```
## [1] TRUE
```

```
print(cell_id_to_enodeb(13067274)==51044)
```

```
## [1] TRUE
```

```
print(is.na(cell_id_to_enodeb(NA)))
```

```
## [1] TRUE
```

```
print(is.na(cell_id_to_enodeb(0)))
```

```
## [1] TRUE
```

Now update the datasets to also contain the eNodeB-ID:

```
dataset_ul = dataset_ul %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_dl = dataset_dl %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_context = dataset_context %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
dataset_cells = dataset_cells %>% mutate(enodeb=map_int(ci, cell_id_to_enodeb))
```

Add an ID for each drive

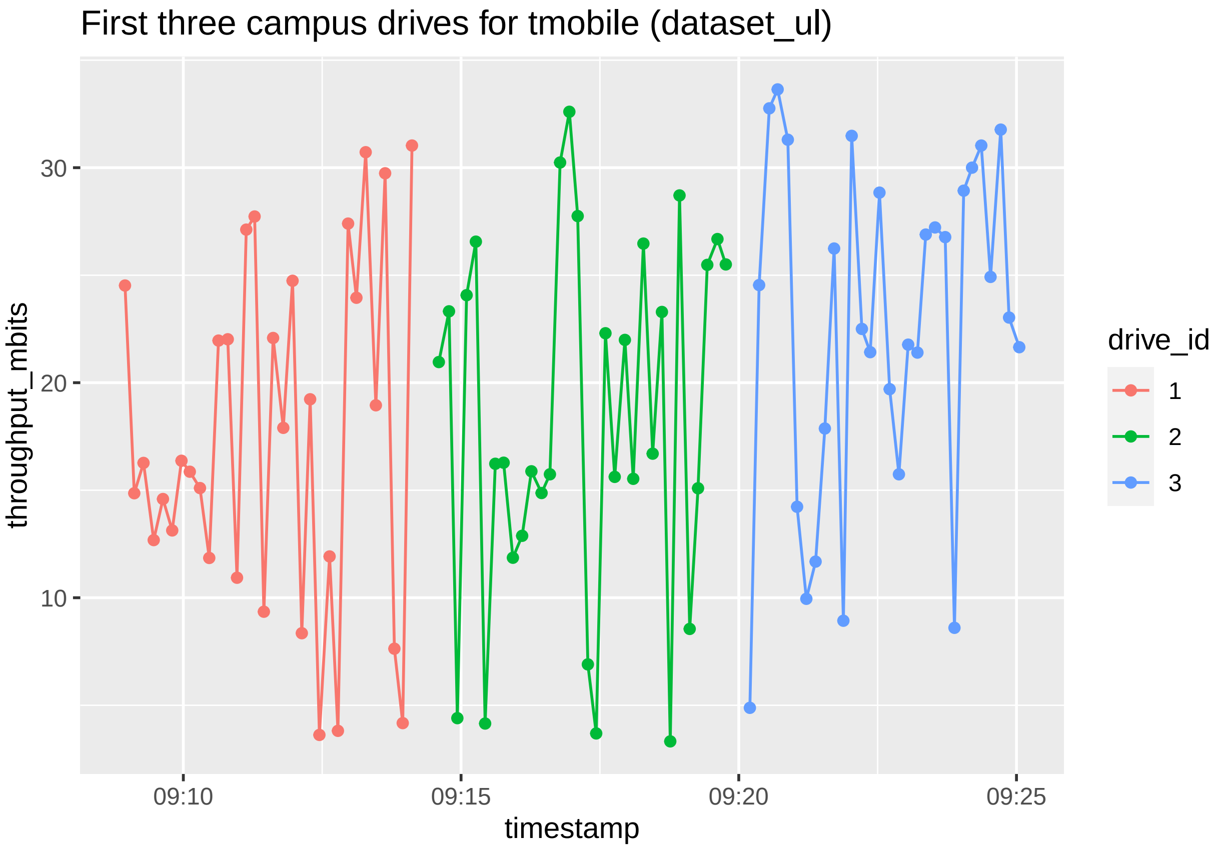
```
get_drive_ids = function(data) {

  num_rows = nrow(data)
  drive_ids = integer(num_rows)

  last_start_time_by_scenario = character(4)
  names(last_start_time_by_scenario) = c("urban", "suburban", "campus", "highway")

  last_drive_id_by_scenario = integer(4)
  names(last_drive_id_by_scenario) =c("urban", "suburban", "campus", "highway")

  for(cur_row in seq_len(num_rows)) {

    cur_scenario = data[[cur_row, "scenario"]]
    cur_start_time = as.character(data[[cur_row, "start_time"]])

    last_start_time = last_start_time_by_scenario[cur_scenario]
    last_drive_id = last_drive_id_by_scenario[cur_scenario]

    if (cur_start_time != last_start_time) {
      cur_drive_id = last_drive_id + 1
      drive_ids[cur_row] = cur_drive_id
      last_start_time_by_scenario[cur_scenario] = cur_start_time
      last_drive_id_by_scenario[cur_scenario] = cur_drive_id
      next()
    }

    drive_ids[cur_row] = last_drive_id
  }

  return(drive_ids)
}
```

```
add_drive_ids = function(dataset) {

  dataset = arrange(dataset, start_time)

  dataset_o2 = filter(dataset, provider=="o2")
  dataset_tmobile = filter(dataset, provider=="tmobile")
  dataset_vodafone = filter(dataset, provider=="vodafone")

  result = bind_rows(
    add_column(dataset_o2, drive_id = factor(get_drive_ids(dataset_o2))),
    add_column(dataset_tmobile, drive_id = factor(get_drive_ids(dataset_tmobile))),
    add_column(dataset_vodafone, drive_id = factor(get_drive_ids(dataset_vodafone)))
  )

  return(result)
}
```

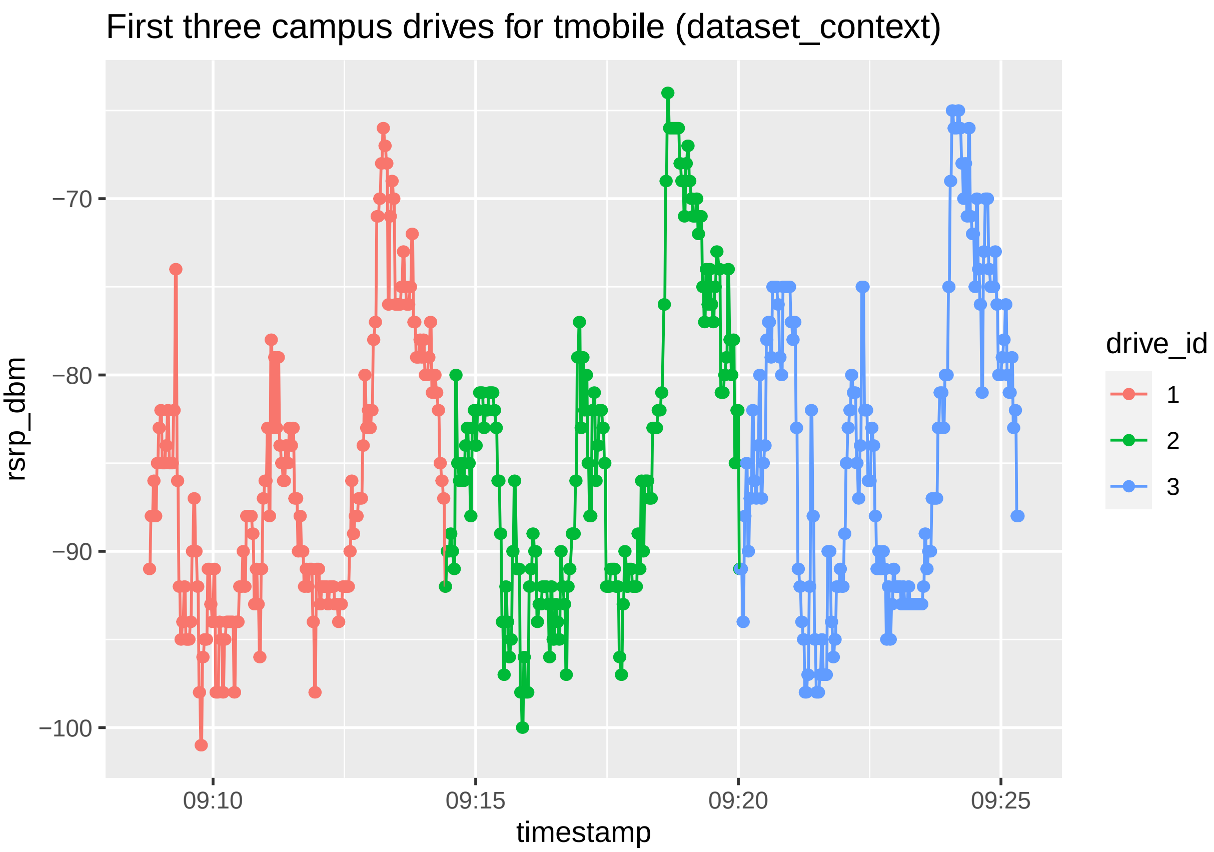
```
dataset_ul = add_drive_ids(dataset_ul)
dataset_dl = add_drive_ids(dataset_dl)
```

```
dataset_context = add_drive_ids(dataset_context)
dataset_cells = add_drive_ids(dataset_cells)
```

```
ggplot(
  filter(dataset_ul, timestamp<=ymd(20181211), provider=="tmobile", scenario=="campus"),
  aes(x=timestamp, y=throughput_mbits)
) + geom_point(aes(color=drive_id)) +
  geom_line(aes(color=drive_id)) +
  ggtitle("First three campus drives for tmobile (dataset_ul)")
```



```
ggplot(
  filter(dataset_context, timestamp<=ymd(20181211), provider=="tmobile", scenario=="campus"),
  aes(x=timestamp, y=rsrp_dbm)
) + geom_point(aes(color=drive_id)) +
  geom_line(aes(color=drive_id)) +
  ggtitle("First three campus drives for tmobile (dataset_context)")
```



Add Neighboring RSRP and RSRQ in Context File

```r
add_neighbor_info = function(dataset_context, dataset_cells) {

  dataset_context <- dataset_context %>% drop_na() %>% arrange(timestamp)
  dataset_cells <- filter(dataset_cells, rsrp_dbm!=0)

  # füge zusätzliche Spalte mit erstmal nur Nullen in context, wo dann am Ende die errechneten
  # RSRP der Nachbarzellen eingefügt werden sollen
  dataset_context$rsrp_neighbor <- 0
  dataset_context$rsrq_neighbor <- 0

  for (p in c("o2", "tmobile", "vodafone")){

    dataset_context_p = filter(dataset_context, provider==p)
    dataset_cells_p <- filter(dataset_cells, provider==p)

    for (s in unique(dataset_context_p$scenario)){

      dataset_context_ps <- filter(dataset_context_p, scenario==s)
        dataset_cells_ps <- filter(dataset_cells_p, scenario==s)

        for (d in unique(dataset_context_ps$drive_id)){
```

```
    dataset_context_psd <- filter(dataset_context_ps, drive_id==d)
      dataset_cells_psd <- filter(dataset_cells_ps, drive_id==d)

    for (i in 1:dim(dataset_context_psd)[1]){

        if (any(dataset_cells_psd$timestamp == dataset_context_psd$timestamp[i])){
          neighbor <- dataset_cells_psd[which(dataset_cells_psd$timestamp == dataset_context_p
        max_rsrp <- max(neighbor$rsrp_dbm)
        dataset_context_psd$rsrp_neighbor[i] <- max_rsrp

          neighbor <- dataset_cells_psd[which(dataset_cells_psd$timestamp == dataset_context_p
        max_rsrq <- max(neighbor$rsrq_db)
        dataset_context_psd$rsrq_neighbor[i] <- max_rsrq
        }
        else {
          if (any(dataset_cells_psd$timestamp < dataset_context_psd$timestamp[i])){
            dataset_context_psd$rsrp_neighbor[i] <- dataset_context_psd$rsrp_neighbor[i-1]
            dataset_context_psd$rsrq_neighbor[i] <- dataset_context_psd$rsrq_neighbor[i-1]
          }
          else {
            dataset_context_psd$rsrp_neighbor[i] <- -Inf
            dataset_context_psd$rsrq_neighbor[i] <- -Inf
          }
        }
    }

      dataset_context[
        (dataset_context$provider == p & dataset_context$scenario == s & dataset_context$drive_id
      ] <- dataset_context_psd
    }
  }
 }

 return(dataset_context)
}
```

```
dataset_context = add_neighbor_info(dataset_context, dataset_cells)
```

Add Link Lifetime to Context File

```
add_link_lifetime = function(dataset_context) {
 dataset_context <- dataset_context %>% drop_na() %>% arrange(timestamp)
 dataset_context$link_lifetime <- -1


 for (p in c("o2", "tmobile", "vodafone")){
  dataset_context_p = filter(dataset_context, provider==p)

  for (s in unique(dataset_context_p$scenario)){
   dataset_context_ps = filter(dataset_context_p, scenario==s)

   for (d in unique(dataset_context_ps$drive_id)){
        dataset_context_psd = filter(dataset_context_ps, drive_id==d)

        for (j in 1:dim(dataset_context_psd)[1]){
          diff_bildungs_ts = dataset_context_psd$time_s[dim(dataset_context_psd)[1]]
          for (i in min(j+1, dim(dataset_context_psd)[1]):dim(dataset_context_psd)[1]){
            if (dataset_context_psd$enodeb[i] != dataset_context_psd$enodeb[j]){
              diff_bildungs_ts <- dataset_context_psd$time_s[i]
              break
            }
          }

          dataset_context_psd$link_lifetime[j] <- diff_bildungs_ts - dataset_context_psd$time_s[j]
        }

        dataset_context[
          (dataset_context$provider == p & dataset_context$scenario == s & dataset_context$drive_id ==
        ] <- dataset_context_psd
      }
    }
  }
  return(dataset_context)
}
```

```
dataset_context = add_link_lifetime(dataset_context)
```

Store the data

```
write_csv(dataset_ul, str_c(data_destination_dir, "dataset_ul.csv"))
write_csv(dataset_dl, str_c(data_destination_dir, "dataset_dl.csv"))
write_csv(dataset_context, str_c(data_destination_dir, "dataset_context.csv"))
write_csv(dataset_cells, str_c(data_destination_dir, "dataset_cells.csv"))
```