Exploratory Data Analysis

Marcelle Chiriboga

2019-05-13

Load the Data

Our data is composed of two tables:

- **Exception Hours** (`exception_hours.csv`): contains the data related to exceptions occurred (or scheduled) between 2012 and 2019.

`train.csv` is the training set and it contains the data related to exceptions logged until 2017. It was created from `exception_hours.csv` by running it through the `src/split_train.R` script.

- **Productive Hours** (`productive_hours.csv`): contains the data related to hours worked from 2010 to present day.

In order to be able to analyze both tables together to compare expections with productive hours, we join both tables to bring the `WORKED_HRS` column into the `exception_hours` table.

```
# Aggregate the exceptions by PROGRAM, COST_CENTRE, JOB_FAMILY_DESCRIPTION, SHIFT_DATE, JOB_STATUS
exception_hours_agg <- exception_hours %>%
  group_by(PROGRAM, COST_CENTRE, JOB_FAMILY_DESCRIPTION,
           SHIFT_DATE, JOB_STATUS) %>%
  summarise(total_exception_hours = sum(EXCEPTION_HOURS),
            number_of_exceptions = n())

# Join tables
exception_and_productive_hours <- prod_hours %>%
  left_join(exception_hours_agg, by = c("PROGRAM", "COST_CENTRE",
                                        "JOB_FAMILY_DESCRIPTION", "SHIFT_DATE",
                                        "FULL_PART_TIME" = "JOB_STATUS")) %>%
  # remove data from 2012, since we don't have exception info for this period
  filter(year(SHIFT_DATE) > 2012)

# Replace NA values with 0
columns <- c("total_exception_hours","number_of_exceptions")
exception_and_productive_hours[columns][is.na(exception_and_productive_hours[columns])] <- 0
```

Exploratory Data Analysis (EDA)

First, lets focus only on the `exception_hours.csv`, exploring how exceptions are distributed accross some of the variables.

Exploring the 'exception_hours' data set

SITE

```
# Check the total number of exceptions by facilities
(facilities <- exception_hours %>%
  group_by(SITE) %>%
  filter(SITE %in% c("Billable", "Brock Fahrni", "Holy Family",
```

```
                      "Mt St Joseph", "PHC Corporate", "St John Hospice",
                      "St Paul's Hospital", "SVH Honoria Conway", "SVH Langara",
                      "Youville Residence")) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
 )
```

```
## # A tibble: 10 x 2
##    SITE               count
##    <chr>              <int>
##  1 St Paul's Hospital 420961
##  2 Mt St Joseph        83590
##  3 Holy Family         37197
##  4 SVH Langara         29193
##  5 PHC Corporate       24002
##  6 Brock Fahrni        19530
##  7 Youville Residence  15678
##  8 SVH Honoria Conway   2799
##  9 St John Hospice      2154
## 10 Billable              555
```

Observation:

- Considering the total number of exceptions from 2013 to 2017, `St Paul's Hospital`, `Mt St Joseph`, `Holy Family` are the top facilities, where `St Paul's Hospital` has ~5x more exceptions than the second facility, `Mt St Joseph`.

We're focusing on the 10 facilities which include `LABOR_AGREEMENT = NURS`. Do we need to include any others?

> As discussed with Sam, there is no need to include other facilities and we won't consider the following sites on our analysis: `PHC Corporate`, `SVH Honoria Conway`, `St John Hospice` and `Billable`. In other words, we'll focus on: `St Paul's Hospital`, `Mt St Joseph`, `Holy Family`, `SVH Langara`, `Brock Fahrni`, `Youville Residence`.

```
facilities <- c("St Paul's Hospital", "Mt St Joseph", "Holy Family", "SVH Langara", "Brock Fahrni", "Yo

# Subset the `exception_hours` dataset in order to just consider the facilities mentioned
exception_hours <- exception_hours %>% filter(SITE %in% facilities)
```

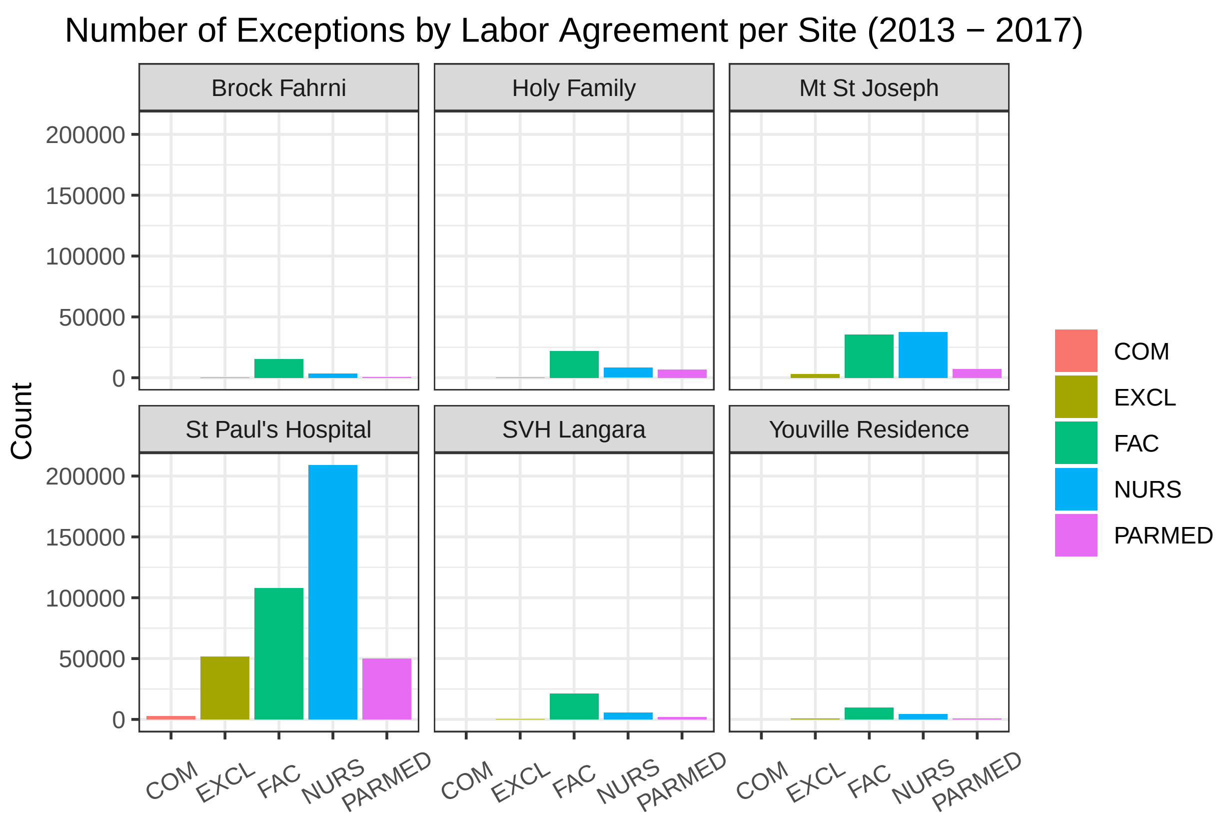
`LABOR_AGREEMENT`

```
# Rank the total number of exceptions by labor agreement
(labor_agreement <- exception_hours %>%
  group_by(LABOR_AGREEMENT) %>%
  filter(!(LABOR_AGREEMENT %in% c('NULL', '0'))) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
)
```

```
## # A tibble: 5 x 2
##   LABOR_AGREEMENT  count
##   <chr>            <int>
## 1 NURS            267786
## 2 FAC             212177
## 3 PARMED           67485
```

```
## 4 EXCL                56047
## 5 COM                  2613
```

```r
# Visualize the total number of exceptions by labor agreement facetting by site
exception_hours %>%
  filter(!(LABOR_AGREEMENT %in% c('NULL', '0')), SITE %in% c("Billable", "Brock Fahrni",
                                                              "Holy Family", "Mt St Joseph",
                                                              "PHC Corporate",
                                                              "St John Hospice",
                                                              "St Paul's Hospital",
                                                              "SVH Honoria Conway",
                                                              "SVH Langara",
                                                              "Youville Residence")) %>%
  ggplot(aes(x = LABOR_AGREEMENT, fill = LABOR_AGREEMENT)) +
  geom_bar(stat = "count") +
  facet_wrap(~SITE) +
  theme_bw() +
  ggtitle("Number of Exceptions by Labor Agreement per Site (2013 - 2017)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 30, hjust = 0.5, vjust = 0.5)) +
  labs(x = "", y = "Count", fill = "")
```



Observations:

- Considering the total number of exceptions from 2013 to 2017, `NURS`, `FAC` and `PARMED` are the top 3 `LABOR_AGREEMENT`.
- Most of the exceptions are from `St. Paul's Hospital`, where the majority are related to `NURS`.

Should we focus only on the top three LABOR_AGREEMENT? Or is there value to analyzing all of the groups, even the less representative ones?

As discussed with Sam, our analysis will focus on NURS, FAC and PARMED LABOR_AGREEMENT.

```
labor_agreements <- c("NURS", "FAC", "PARMED")

# Subset the `exception_hours` dataset in order to just consider the top 3 LABOR_AGREEMENT
exception_hours <- exception_hours %>% filter(LABOR_AGREEMENT %in% LABOR_AGREEMENT)
```

JOB_FAMILY_DESCRIPTION

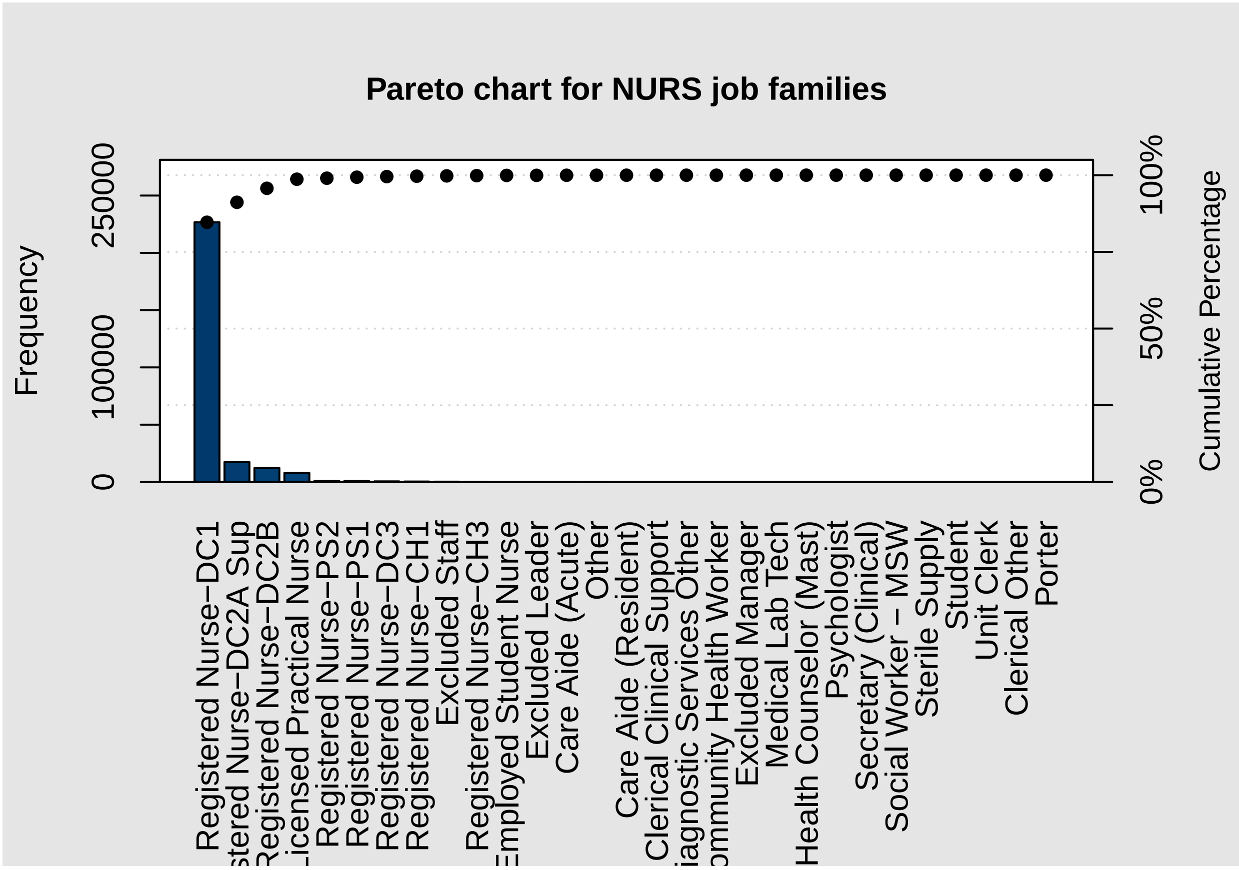
Exploring the JOB_FAMILY_DESCRIPTION of the main LABOR_AGREEMENT:

- NURS job families (top 10)

```
# NURS job families
(nurs_job_family <- exception_hours %>%
  filter(LABOR_AGREEMENT == "NURS") %>%
  group_by(JOB_FAMILY_DESCRIPTION) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(cumsum = cumsum(count),
         freq = round(count / sum(count), 3),
         cum_freq = cumsum(freq))
)
```

```
## # A tibble: 29 x 5
##    JOB_FAMILY_DESCRIPTION     count cumsum  freq cum_freq
##    <chr>                      <int>  <int> <dbl>    <dbl>
##  1 Registered Nurse-DC1      226688 226688 0.847    0.847
##  2 Registered Nurse-DC2A Sup  17433 244121 0.065    0.912
##  3 Registered Nurse-DC2B      12216 256337 0.046    0.958
##  4 Licensed Practical Nurse    7925 264262 0.03     0.988
##  5 Registered Nurse-PS2         890 265152 0.003    0.991
##  6 Registered Nurse-PS1         861 266013 0.003    0.994
##  7 Registered Nurse-DC3         521 266534 0.002    0.996
##  8 Registered Nurse-CH1         368 266902 0.001    0.997
##  9 Excluded Staff               266 267168 0.001    0.998
## 10 Registered Nurse-CH3         214 267382 0.001    0.999
## # ... with 19 more rows
```

```
# Plot pareto
nurs_job_family_count <- nurs_job_family$count
names(nurs_job_family_count) <- nurs_job_family$`JOB_FAMILY_DESCRIPTION`
pareto.chart(nurs_job_family_count, cumperc = seq(0, 100, by = 25), main = "Pareto chart for NURS job fa
```




```
##
## Pareto chart analysis for nurs_job_family_count
##                                    Frequency    Cum.Freq.   Percentage
##   Registered Nurse-DC1           2.266880e+05 2.266880e+05 8.465267e+01
##   Registered Nurse-DC2A Sup      1.743300e+04 2.441210e+05 6.510049e+00
##   Registered Nurse-DC2B          1.221600e+04 2.563370e+05 4.561852e+00
##   Licensed Practical Nurse       7.925000e+03 2.642620e+05 2.959453e+00
##   Registered Nurse-PS2           8.900000e+02 2.651520e+05 3.323549e-01
##   Registered Nurse-PS1           8.610000e+02 2.660130e+05 3.215254e-01
##   Registered Nurse-DC3           5.210000e+02 2.665340e+05 1.945583e-01
##   Registered Nurse-CH1           3.680000e+02 2.669020e+05 1.374232e-01
##   Excluded Staff                 2.660000e+02 2.671680e+05 9.933305e-02
##   Registered Nurse-CH3           2.140000e+02 2.673820e+05 7.991456e-02
##   Employed Student Nurse         1.790000e+02 2.675610e+05 6.684442e-02
##   Excluded Leader                7.400000e+01 2.676350e+05 2.763401e-02
##   Care Aide (Acute)              6.000000e+01 2.676950e+05 2.240595e-02
##   Other                          4.300000e+01 2.677380e+05 1.605760e-02
##   Care Aide (Resident)           9.000000e+00 2.677470e+05 3.360893e-03
##   Clerical Clinical Support      6.000000e+00 2.677530e+05 2.240595e-03
##   Diagnostic Services Other      5.000000e+00 2.677580e+05 1.867163e-03
##   Community Health Worker        3.000000e+00 2.677610e+05 1.120298e-03
##   Excluded Manager               3.000000e+00 2.677640e+05 1.120298e-03
##   Medical Lab Tech               3.000000e+00 2.677670e+05 1.120298e-03
##   Mental Health Counselor (Mast) 3.000000e+00 2.677700e+05 1.120298e-03
##   Psychologist                   3.000000e+00 2.677730e+05 1.120298e-03
##   Secretary (Clinical)           3.000000e+00 2.677760e+05 1.120298e-03
```

```
##    Social Worker - MSW            2.000000e+00 2.677780e+05 7.468650e-04
##    Sterile Supply                 2.000000e+00 2.677800e+05 7.468650e-04
##    Student                        2.000000e+00 2.677820e+05 7.468650e-04
##    Unit Clerk                     2.000000e+00 2.677840e+05 7.468650e-04
##    Clerical Other                 1.000000e+00 2.677850e+05 3.734325e-04
##    Porter                         1.000000e+00 2.677860e+05 3.734325e-04
##
## Pareto chart analysis for nurs_job_family_count
##                                  Cum.Percent.
##    Registered Nurse-DC1           8.465267e+01
##    Registered Nurse-DC2A Sup      9.116272e+01
##    Registered Nurse-DC2B          9.572457e+01
##    Licensed Practical Nurse       9.868402e+01
##    Registered Nurse-PS2           9.901638e+01
##    Registered Nurse-PS1           9.933790e+01
##    Registered Nurse-DC3           9.953246e+01
##    Registered Nurse-CH1           9.966989e+01
##    Excluded Staff                 9.976922e+01
##    Registered Nurse-CH3           9.984913e+01
##    Employed Student Nurse         9.991598e+01
##    Excluded Leader                9.994361e+01
##    Care Aide (Acute)              9.996602e+01
##    Other                          9.998208e+01
##    Care Aide (Resident)           9.998544e+01
##    Clerical Clinical Support      9.998768e+01
##    Diagnostic Services Other      9.998954e+01
##    Community Health Worker        9.999066e+01
##    Excluded Manager               9.999178e+01
##    Medical Lab Tech               9.999290e+01
##    Mental Health Counselor (Mast) 9.999403e+01
##    Psychologist                   9.999515e+01
##    Secretary (Clinical)           9.999627e+01
##    Social Worker - MSW            9.999701e+01
##    Sterile Supply                 9.999776e+01
##    Student                        9.999851e+01
##    Unit Clerk                     9.999925e+01
##    Clerical Other                 9.999963e+01
##    Porter                         1.000000e+02
```

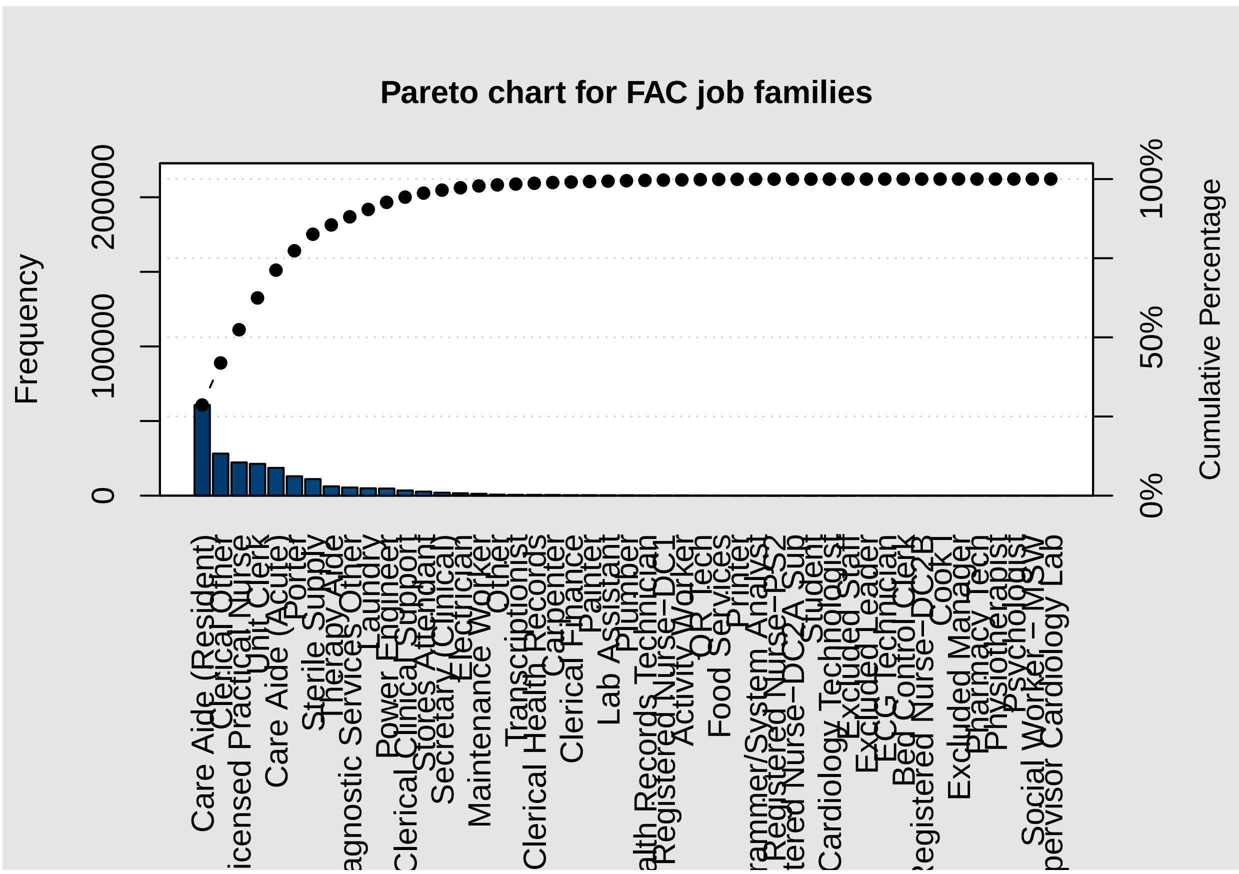
- FAC job families (top 10)

```r
# FAC job families
(fac_job_family <- exception_hours %>%
   filter(LABOR_AGREEMENT == "FAC") %>%
   group_by(JOB_FAMILY_DESCRIPTION) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(cumsum = cumsum(count),
         freq = round(count / sum(count), 3),
         cum_freq = cumsum(freq))
)
```

```
## # A tibble: 47 x 5
##     JOB_FAMILY_DESCRIPTION    count cumsum  freq cum_freq
##     <chr>                     <int>  <int> <dbl>    <dbl>
```

```
##  1 Care Aide (Resident)      60735  60735 0.286    0.286
##  2 Clerical Other            28179  88914 0.133    0.419
##  3 Licensed Practical Nurse  22278 111192 0.105    0.524
##  4 Unit Clerk                21313 132505 0.1      0.624
##  5 Care Aide (Acute)         18601 151106 0.088    0.712
##  6 Porter                    12998 164104 0.061    0.773
##  7 Sterile Supply            11097 175201 0.052    0.825
##  8 Therapy Aide               6228 181429 0.029    0.854
##  9 Diagnostic Services Other  5451 186880 0.026    0.88
## 10 Laundry                    4934 191814 0.023    0.903
## # ... with 37 more rows
```

```r
# Plot pareto
fac_job_family_count <- fac_job_family$count
names(fac_job_family_count) <- fac_job_family$`JOB_FAMILY_DESCRIPTION`
pareto.chart(fac_job_family_count, cumperc = seq(0, 100, by = 25), main = "Pareto chart for FAC job fam
```




```
##
## Pareto chart analysis for fac_job_family_count
##                               Frequency    Cum.Freq.   Percentage
##   Care Aide (Resident)       6.073500e+04 6.073500e+04 2.862469e+01
##   Clerical Other             2.817900e+04 8.891400e+04 1.328089e+01
##   Licensed Practical Nurse   2.227800e+04 1.111920e+05 1.049972e+01
##   Unit Clerk                 2.131300e+04 1.325050e+05 1.004492e+01
##   Care Aide (Acute)          1.860100e+04 1.511060e+05 8.766737e+00
##   Porter                     1.299800e+04 1.641040e+05 6.126017e+00
##   Sterile Supply             1.109700e+04 1.752010e+05 5.230067e+00
```

```
##   Therapy Aide              6.228000e+03 1.814290e+05 2.935285e+00
##   Diagnostic Services Other 5.451000e+03 1.868800e+05 2.569081e+00
##   Laundry                   4.934000e+03 1.918140e+05 2.325417e+00
##   Power Engineer            4.773000e+03 1.965870e+05 2.249537e+00
##   Clerical Clinical Support 3.491000e+03 2.000780e+05 1.645324e+00
##   Stores Attendant          2.697000e+03 2.027750e+05 1.271109e+00
##   Secretary (Clinical)      1.973000e+03 2.047480e+05 9.298840e-01
##   Electrician               1.592000e+03 2.063400e+05 7.503170e-01
##   Maintenance Worker        1.232000e+03 2.075720e+05 5.806473e-01
##   Other                     7.050000e+02 2.082770e+05 3.322698e-01
##   Transcriptionist          5.480000e+02 2.088250e+05 2.582749e-01
##   Clerical Health Records   5.190000e+02 2.093440e+05 2.446071e-01
##   Carpenter                 4.790000e+02 2.098230e+05 2.257549e-01
##   Clerical Finance          3.610000e+02 2.101840e+05 1.701410e-01
##   Painter                   3.480000e+02 2.105320e+05 1.640140e-01
##   Lab Assistant             2.850000e+02 2.108170e+05 1.343218e-01
##   Plumber                   2.660000e+02 2.110830e+05 1.253670e-01
##   Health Records Technician 2.210000e+02 2.113040e+05 1.041583e-01
##   Registered Nurse-DC1      1.980000e+02 2.115020e+05 9.331831e-02
##   Activity Worker           1.720000e+02 2.116740e+05 8.106439e-02
##   OR Tech                   1.370000e+02 2.118110e+05 6.456873e-02
##   Food Services             1.240000e+02 2.119350e+05 5.844177e-02
##   Printer                   7.900000e+01 2.120140e+05 3.723306e-02
##   Programmer/System Analyst 5.500000e+01 2.120690e+05 2.592175e-02
##   Registered Nurse-PS2      2.100000e+01 2.120900e+05 9.897397e-03
##   Registered Nurse-DC2A Sup 1.900000e+01 2.121090e+05 8.954788e-03
##   Student                   1.900000e+01 2.121280e+05 8.954788e-03
##   Cardiology Technologist   1.300000e+01 2.121410e+05 6.126960e-03
##   Excluded Staff            1.000000e+01 2.121510e+05 4.713046e-03
##   Excluded Leader           6.000000e+00 2.121570e+05 2.827828e-03
##   ECG Technician            5.000000e+00 2.121620e+05 2.356523e-03
##   Bed Control Clerk         4.000000e+00 2.121660e+05 1.885218e-03
##   Registered Nurse-DC2B     4.000000e+00 2.121700e+05 1.885218e-03
##   Cook I                    1.000000e+00 2.121710e+05 4.713046e-04
##   Excluded Manager          1.000000e+00 2.121720e+05 4.713046e-04
##   Pharmacy Tech             1.000000e+00 2.121730e+05 4.713046e-04
##   Physiotherapist           1.000000e+00 2.121740e+05 4.713046e-04
##   Psychologist              1.000000e+00 2.121750e+05 4.713046e-04
##   Social Worker - MSW       1.000000e+00 2.121760e+05 4.713046e-04
##   Supervisor Cardiology Lab 1.000000e+00 2.121770e+05 4.713046e-04
##
## Pareto chart analysis for fac_job_family_count
##                             Cum.Percent.
##   Care Aide (Resident)      2.862469e+01
##   Clerical Other            4.190558e+01
##   Licensed Practical Nurse  5.240530e+01
##   Unit Clerk                6.245022e+01
##   Care Aide (Acute)         7.121696e+01
##   Porter                    7.734297e+01
##   Sterile Supply            8.257304e+01
##   Therapy Aide              8.550833e+01
##   Diagnostic Services Other 8.807741e+01
##   Laundry                   9.040282e+01
##   Power Engineer            9.265236e+01
```

```
##   Clerical Clinical Support 9.429769e+01
##   Stores Attendant          9.556879e+01
##   Secretary (Clinical)      9.649868e+01
##   Electrician               9.724899e+01
##   Maintenance Worker        9.782964e+01
##   Other                     9.816191e+01
##   Transcriptionist          9.842019e+01
##   Clerical Health Records   9.866479e+01
##   Carpenter                 9.889055e+01
##   Clerical Finance          9.906069e+01
##   Painter                   9.922470e+01
##   Lab Assistant             9.935903e+01
##   Plumber                   9.948439e+01
##   Health Records Technician 9.958855e+01
##   Registered Nurse-DC1      9.968187e+01
##   Activity Worker           9.976293e+01
##   OR Tech                   9.982750e+01
##   Food Services             9.988594e+01
##   Printer                   9.992318e+01
##   Programmer/System Analyst 9.994910e+01
##   Registered Nurse-PS2      9.995900e+01
##   Registered Nurse-DC2A Sup 9.996795e+01
##   Student                   9.997691e+01
##   Cardiology Technologist   9.998303e+01
##   Excluded Staff            9.998775e+01
##   Excluded Leader           9.999057e+01
##   ECG Technician            9.999293e+01
##   Bed Control Clerk         9.999482e+01
##   Registered Nurse-DC2B     9.999670e+01
##   Cook I                    9.999717e+01
##   Excluded Manager          9.999764e+01
##   Pharmacy Tech             9.999811e+01
##   Physiotherapist           9.999859e+01
##   Psychologist              9.999906e+01
##   Social Worker - MSW       9.999953e+01
##   Supervisor Cardiology Lab 1.000000e+02
```

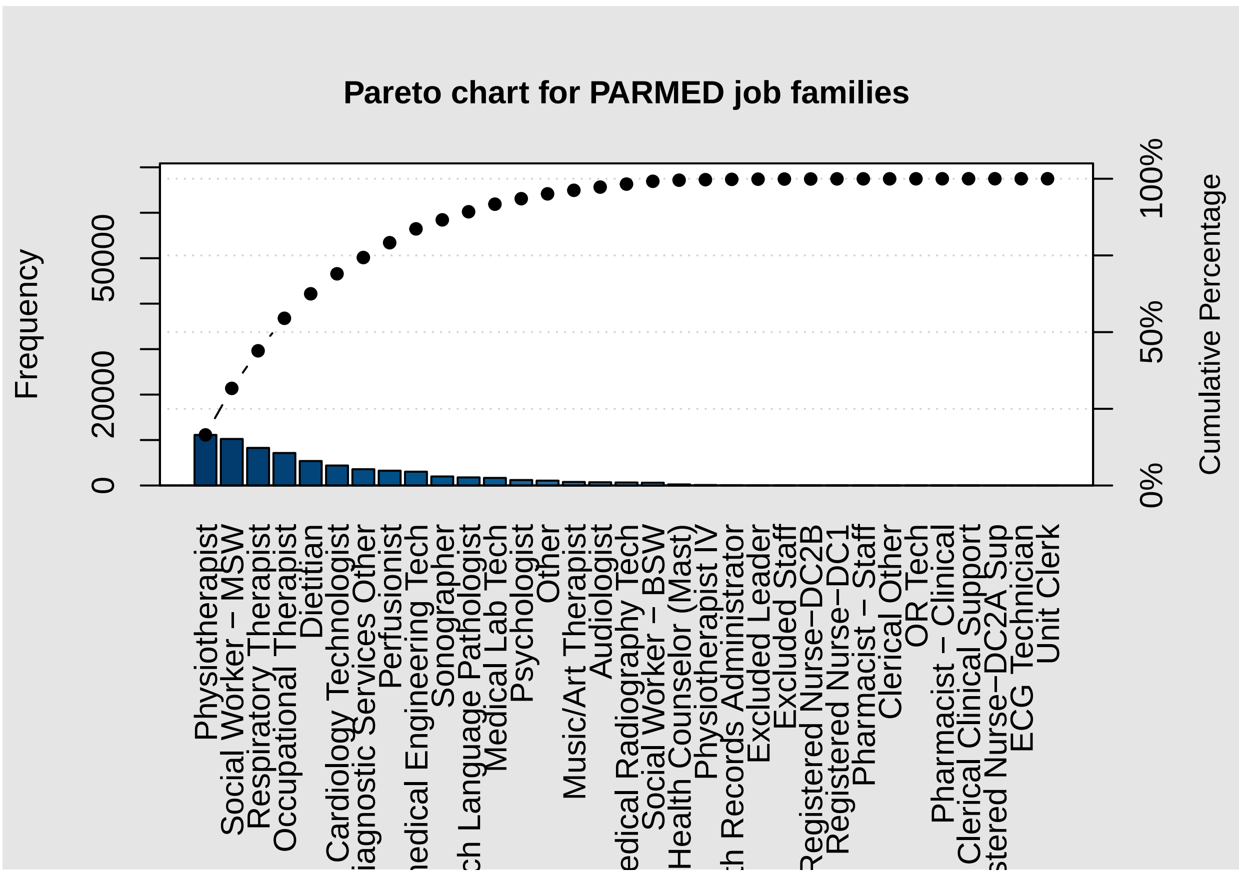
- PARMED job families (top 10)

```
# PARMED job families
(parmed_job_family <- exception_hours %>%
   filter(LABOR_AGREEMENT == "PARMED") %>%
   group_by(JOB_FAMILY_DESCRIPTION) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(cumsum = cumsum(count),
        freq = round(count / sum(count), 3),
        cum_freq = cumsum(freq))
)
```

```
## # A tibble: 33 x 5
##    JOB_FAMILY_DESCRIPTION      count cumsum  freq cum_freq
##    <chr>                       <int>  <int> <dbl>    <dbl>
##  1 Physiotherapist             11126  11126 0.165    0.165
##  2 Social Worker - MSW         10236  21362 0.152    0.317
```

```
##   3 Respiratory Therapist        8271  29633 0.123    0.44
##   4 Occupational Therapist       7152  36785 0.106    0.546
##   5 Dietitian                    5387  42172 0.08     0.626
##   6 Cardiology Technologist      4397  46569 0.065    0.691
##   7 Diagnostic Services Other    3583  50152 0.053    0.744
##   8 Perfusionist                 3260  53412 0.048    0.792
##   9 Biomedical Engineering Tech  3040  56452 0.045    0.837
##  10 Sonographer                  1985  58437 0.029    0.866
## # ... with 23 more rows
```

```
# Plot pareto
parmed_job_family_count <- parmed_job_family$count
names(parmed_job_family_count) <- parmed_job_family$`JOB_FAMILY_DESCRIPTION`
pareto.chart(parmed_job_family_count, cumperc = seq(0, 100, by = 25), main = "Pareto chart for PARMED j
```




```
##
## Pareto chart analysis for parmed_job_family_count
##                                 Frequency     Cum.Freq.   Percentage
##    Physiotherapist             1.112600e+04 1.112600e+04 1.648663e+01
##    Social Worker - MSW         1.023600e+04 2.136200e+04 1.516782e+01
##    Respiratory Therapist       8.271000e+03 2.963300e+04 1.225606e+01
##    Occupational Therapist      7.152000e+03 3.678500e+04 1.059791e+01
##    Dietitian                   5.387000e+03 4.217200e+04 7.982515e+00
##    Cardiology Technologist     4.397000e+03 4.656900e+04 6.515522e+00
##    Diagnostic Services Other   3.583000e+03 5.015200e+04 5.309328e+00
##    Perfusionist                3.260000e+03 5.341200e+04 4.830703e+00
##    Biomedical Engineering Tech 3.040000e+03 5.645200e+04 4.504705e+00
```

```
##   Sonographer                      1.985000e+03 5.843700e+04 2.941394e+00
##   Speech Language Pathologist      1.782000e+03 6.021900e+04 2.640587e+00
##   Medical Lab Tech                 1.671000e+03 6.189000e+04 2.476106e+00
##   Psychologist                     1.199000e+03 6.308900e+04 1.776691e+00
##   Other                            1.071000e+03 6.416000e+04 1.587019e+00
##   Music/Art Therapist              7.790000e+02 6.493900e+04 1.154331e+00
##   Audiologist                      7.070000e+02 6.564600e+04 1.047640e+00
##   Medical Radiography Tech         6.600000e+02 6.630600e+04 9.779951e-01
##   Social Worker - BSW              6.160000e+02 6.692200e+04 9.127954e-01
##   Mental Health Counselor (Mast) 2.420000e+02 6.716400e+04 3.585982e-01
##   Physiotherapist IV               1.110000e+02 6.727500e+04 1.644810e-01
##   Health Records Administrator     7.500000e+01 6.735000e+04 1.111358e-01
##   Excluded Leader                  3.800000e+01 6.738800e+04 5.630881e-02
##   Excluded Staff                   2.200000e+01 6.741000e+04 3.259984e-02
##   Registered Nurse-DC2B            2.100000e+01 6.743100e+04 3.111803e-02
##   Registered Nurse-DC1             1.900000e+01 6.745000e+04 2.815440e-02
##   Pharmacist - Staff               9.000000e+00 6.745900e+04 1.333630e-02
##   Clerical Other                   7.000000e+00 6.746600e+04 1.037268e-02
##   OR Tech                          6.000000e+00 6.747200e+04 8.890865e-03
##   Pharmacist - Clinical            5.000000e+00 6.747700e+04 7.409054e-03
##   Clerical Clinical Support        4.000000e+00 6.748100e+04 5.927243e-03
##   Registered Nurse-DC2A Sup        2.000000e+00 6.748300e+04 2.963622e-03
##   ECG Technician                   1.000000e+00 6.748400e+04 1.481811e-03
##   Unit Clerk                       1.000000e+00 6.748500e+04 1.481811e-03
##
## Pareto chart analysis for parmed_job_family_count
##                                    Cum.Percent.
##   Physiotherapist                  1.648663e+01
##   Social Worker - MSW              3.165444e+01
##   Respiratory Therapist            4.391050e+01
##   Occupational Therapist           5.450841e+01
##   Dietitian                        6.249092e+01
##   Cardiology Technologist          6.900645e+01
##   Diagnostic Services Other        7.431577e+01
##   Perfusionist                     7.914648e+01
##   Biomedical Engineering Tech      8.365118e+01
##   Sonographer                      8.659258e+01
##   Speech Language Pathologist      8.923316e+01
##   Medical Lab Tech                 9.170927e+01
##   Psychologist                     9.348596e+01
##   Other                            9.507298e+01
##   Music/Art Therapist              9.622731e+01
##   Audiologist                      9.727495e+01
##   Medical Radiography Tech         9.825295e+01
##   Social Worker - BSW              9.916574e+01
##   Mental Health Counselor (Mast) 9.952434e+01
##   Physiotherapist IV               9.968882e+01
##   Health Records Administrator     9.979996e+01
##   Excluded Leader                  9.985626e+01
##   Excluded Staff                   9.988886e+01
##   Registered Nurse-DC2B            9.991998e+01
##   Registered Nurse-DC1             9.994814e+01
##   Pharmacist - Staff               9.996147e+01
##   Clerical Other                   9.997185e+01
```

```
##    OR Tech                          9.998074e+01
##    Pharmacist - Clinical            9.998815e+01
##    Clerical Clinical Support        9.999407e+01
##    Registered Nurse-DC2A Sup        9.999704e+01
##    ECG Technician                   9.999852e+01
##    Unit Clerk                       1.000000e+02
```

EXCEPTION_GROUP

- Considering all sites

```r
# Check the total number of exceptions by each exception group
(exception_groups <- exception_hours %>%
  group_by(EXCEPTION_GROUP) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
 )
```

```
## # A tibble: 15 x 2
##    EXCEPTION_GROUP           count
##    <chr>                     <int>
##  1 Other                    164408
##  2 Vacation                  96428
##  3 Swap                      76853
##  4 Paid Sick                 60181
##  5 Workload                  58924
##  6 Vacancy                   43215
##  7 Leave of Absence          32071
##  8 Move                      17673
##  9 Casual Sick or Cancelled  16331
## 10 Unpaid Sick                9127
## 11 Work Related Injury        8625
## 12 Schedule Adjustment        7504
## 13 Education                  6354
## 14 On Call & Call Back        5756
## 15 Relief Sick                2699
```

Most exceptions fall under`Other`. Let's look at those to see what are some of the exception reasons associated under this group.

```r
# Check the total number of `Other` exceptions by each exception reason
other_exception_reason <- exception_hours %>%
  filter(EXCEPTION_GROUP == "Other") %>%
  group_by(EXCEPTION_REASON) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

head(other_exception_reason, 10)
```

```
## # A tibble: 10 x 2
##    EXCEPTION_REASON                  count
##    <chr>                             <int>
##  1 PVC- Vacation Regular - MV- Move 26566
##  2 REG- Regular Hrs - MV- Move      21812
##  3 FTE- Flex Time Earned NC         19559
##  4 REG- Regular Hrs                 14160
```

```
##   5 Vacant Shift - MV- Move          12508
##   6 OGX- OT Meeting 1x               11094
##   7 PSK- Sick Lv - MV- Move           5424
##   8 REG- Wkld Increase - MV- Move     5156
##   9 Swap shifts - MV- Move            3485
##  10 BGX- OT Bank Meeting 1x           2842
```

- Focusing on `St Paul's Hospital` `EXCEPTION_GROUP` to check if the main groups are the same as the ones considering PHC as a whole.

```r
# Check the St Pauls Hospital total number of exceptions by each exception group
(exception_groups_st_paul <- exception_hours %>%
  filter(SITE == "St Paul's Hospital") %>%
  group_by(EXCEPTION_GROUP) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
)
```

```
## # A tibble: 15 x 2
##    EXCEPTION_GROUP            count
##    <chr>                      <int>
##  1 Other                     117314
##  2 Vacation                   69348
##  3 Swap                       46102
##  4 Workload                   42917
##  5 Paid Sick                  39606
##  6 Vacancy                    32106
##  7 Leave of Absence           22229
##  8 Move                       14271
##  9 Casual Sick or Cancelled   10103
## 10 Unpaid Sick                 6444
## 11 Schedule Adjustment         5503
## 12 On Call & Call Back         4651
## 13 Education                   4419
## 14 Work Related Injury         3981
## 15 Relief Sick                 1967
```

```r
# Check the St Paul's Hospital total number of `Other` exceptions by each exception reason
other_exception_reason_st_paul <- exception_hours %>%
  filter(SITE == "St Paul's Hospital" & EXCEPTION_GROUP == "Other") %>%
  group_by(EXCEPTION_REASON) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

head(other_exception_reason_st_paul, 10)
```

```
## # A tibble: 10 x 2
##    EXCEPTION_REASON                    count
##    <chr>                               <int>
##  1 FTE- Flex Time Earned NC            17428
##  2 REG- Regular Hrs - MV- Move         16691
##  3 PVC- Vacation Regular - MV- Move 14307
##  4 REG- Regular Hrs                    12837
##  5 Vacant Shift - MV- Move              8947
##  6 OGX- OT Meeting 1x                   7275
##  7 REG- Wkld Increase - MV- Move        3603
```

```
##  8 PSK- Sick Lv - MV- Move          3052
##  9 BGX- OT Bank Meeting 1x          2315
## 10 REG- Working Off Site            1964
```

Observations:

Top `EXCEPTION_GROUP` by number of exceptions:

- PHC as a whole: `Other` > `Vacation` > `Paid Sick` > `Swap`
- `St. Paul's Hospital`: `Other` > `Vacation` > `Swap` > `Workload`

Top `EXCEPTION_REASON` related to `Other` `EXCEPTION_GROUP`:

- PHC as a whole: `REG- Regular Hrs - MV- Move`, `PVC- Vacation Regular - MV- Move`, `REG- Regular Hrs`, `FTE- Flex Time Earned NC`
- `St. Paul's Hospital`: `FTE- Flex Time Earned NC`, `REG- Regular Hrs - MV- Move`, `PVC- Vacation Regular - MV- Move`, `REG- Regular Hrs`

Given `Other` is the top 1 `EXCEPTION_GROUP`, and since the `EXCEPTION_REASON` associated seem to, in several cases, fit into one or more of the other existing `EXCEPTION_GROUP`, should we attempt to recategorize some of these exceptions?

St Paul's Hospital - Vacation

Analyze `Vacation` and Sickness ('Paid Sick', 'Unpaid Sick', 'Relief Sick') `EXCEPTION_GROUP` from `St Paul's Hospital`.

```r
# Create a dataset for St Paul's Hospital Vacation
(vacation_weekly <- exception_hours %>%
  filter(SITE == "St Paul's Hospital", EXCEPTION_GROUP == 'Vacation') %>%
  # extract year and week
  mutate(year = year(SHIFT_DATE),
         week = week(SHIFT_DATE)) %>%
  group_by(year, week) %>%
  summarise(count = n()) %>%
  # remove the last week of each year (week 53), since they consider few days
  filter(week != 53)
)
```

```
## # A tibble: 260 x 3
## # Groups:   year [5]
##     year  week count
##    <dbl> <dbl> <int>
##  1  2013     1   296
##  2  2013     2   172
##  3  2013     3   134
##  4  2013     4   126
##  5  2013     5   137
##  6  2013     6   142
##  7  2013     7   198
##  8  2013     8   174
##  9  2013     9   128
## 10  2013    10   150
## # ... with 250 more rows
```

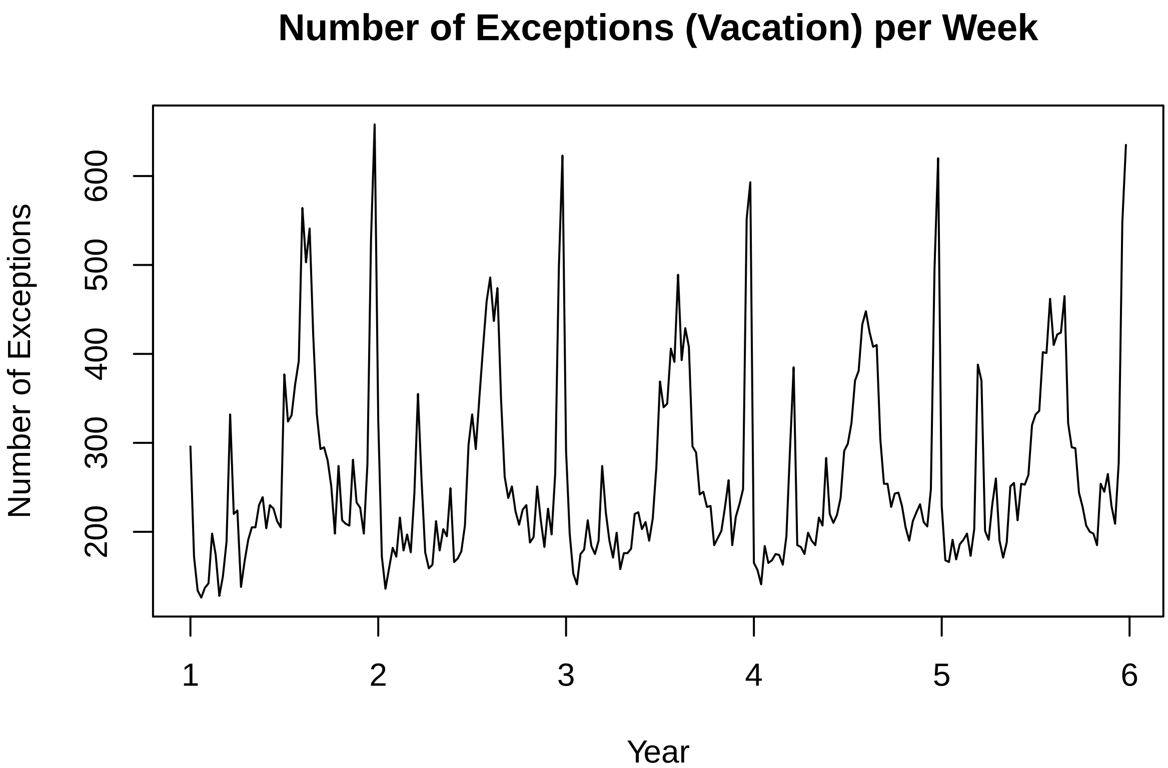
```r
# Create a dataset for sick
(sick_weekly <- exception_hours %>%
  filter(SITE == "St Paul's Hospital",
```

```r
         EXCEPTION_GROUP %in% c('Paid Sick', 'Unpaid Sick', 'Relief Sick')) %>%
    # extract year and week
  mutate(year = year(SHIFT_DATE),
         week = week(SHIFT_DATE)) %>%
  group_by(year, week) %>%
  summarise(count = n()) %>%
  # remove the last week of each year (week 53), since they consider few days
  filter(week != 53)
)
```

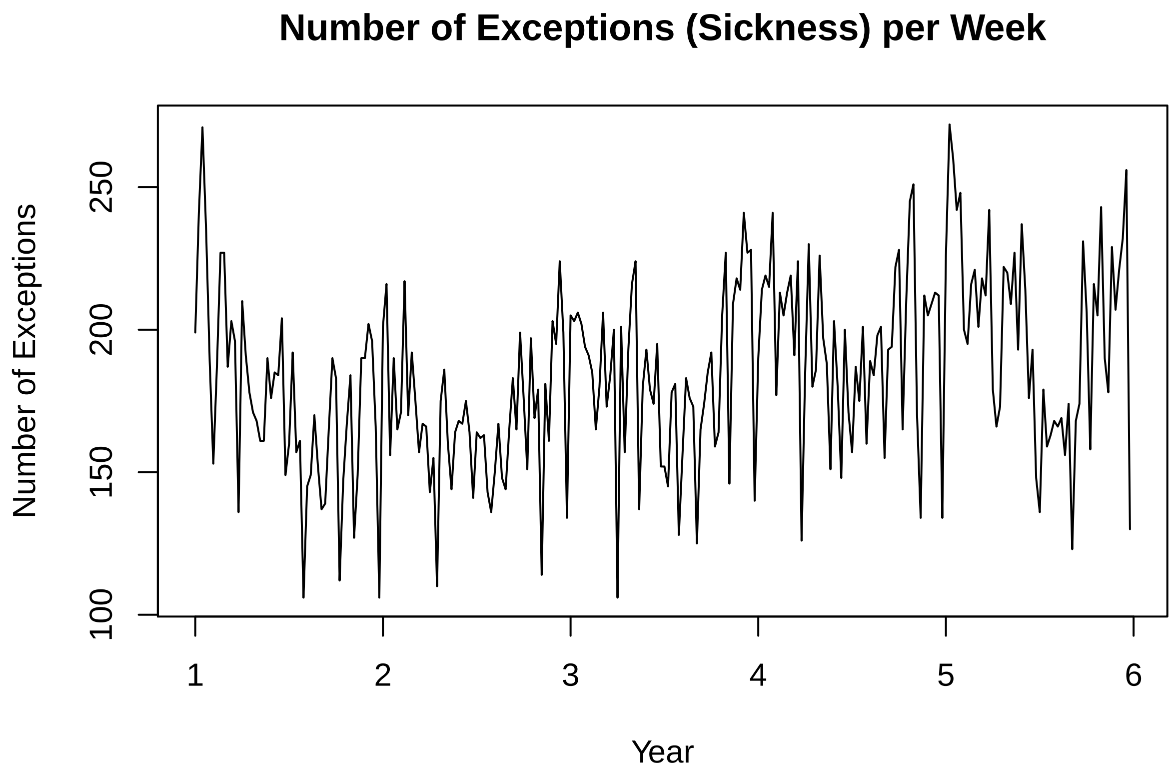
```
## # A tibble: 260 x 3
## # Groups:   year [5]
##     year  week count
##    <dbl> <dbl> <int>
##  1  2013     1   199
##  2  2013     2   241
##  3  2013     3   271
##  4  2013     4   235
##  5  2013     5   189
##  6  2013     6   153
##  7  2013     7   187
##  8  2013     8   227
##  9  2013     9   227
## 10  2013    10   187
## # ... with 250 more rows
```

```r
# Create daily time series for different exception groups
ts_vacation_weekly <- ts(vacation_weekly$count, frequency = 52)
ts_sick_weekly <- ts(sick_weekly$count, frequency = 52)

# Plot the time series
plot(ts_vacation_weekly, xlab = "Year", ylab = "Number of Exceptions",
     main = "Number of Exceptions (Vacation) per Week")
```

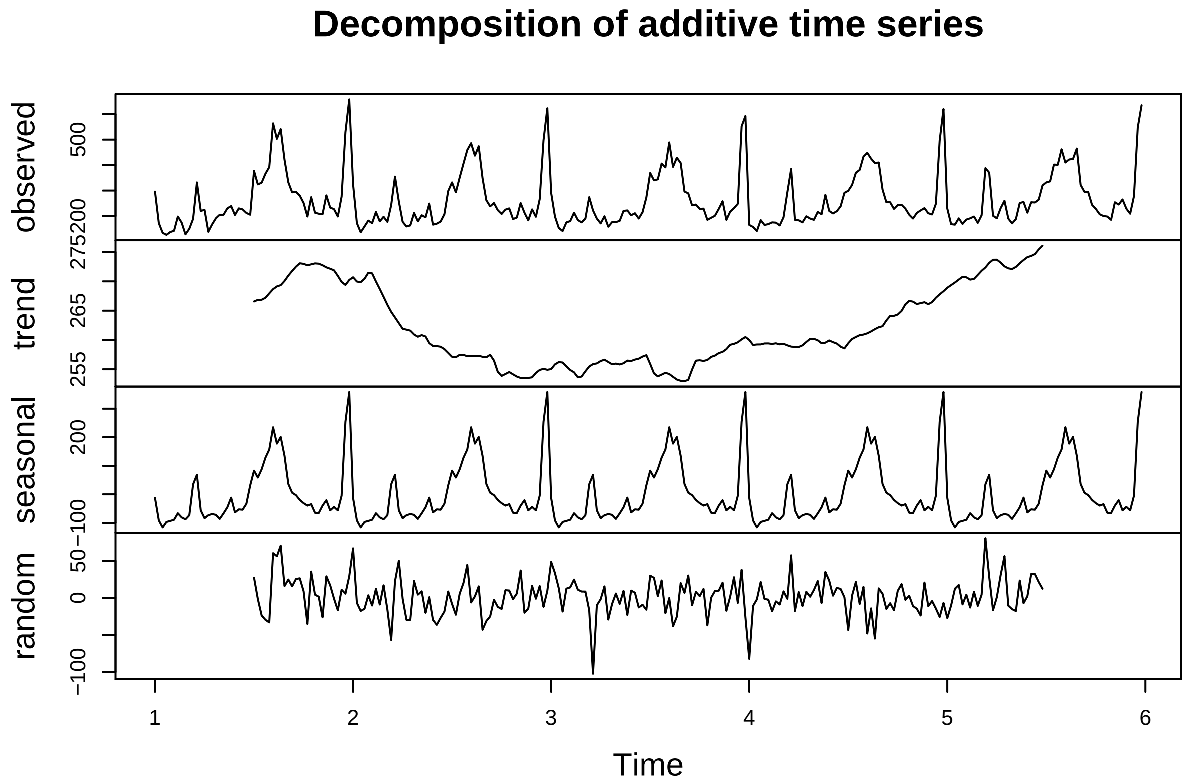



```
plot(ts_sick_weekly, xlab = "Year", ylab = "Number of Exceptions",
     main = "Number of Exceptions (Sickness) per Week")
```

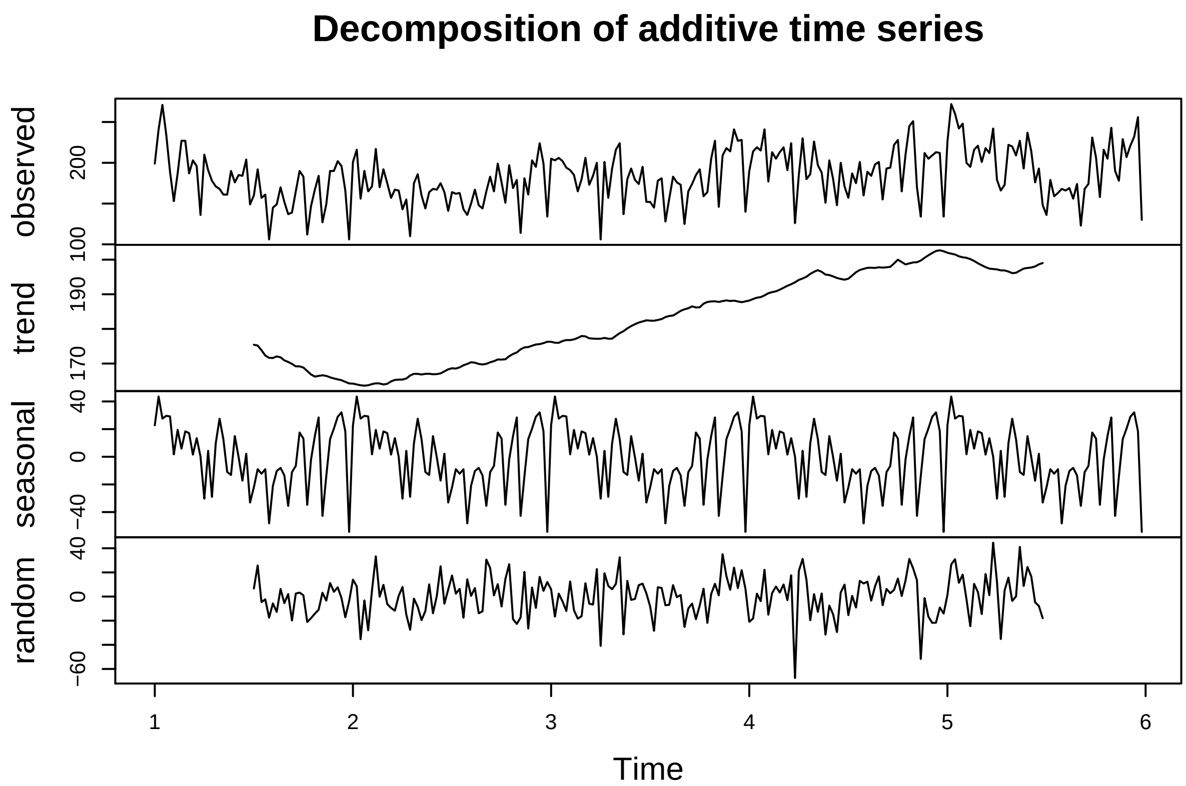



```r
dec_ts_vacation_weekly <- decompose(ts_vacation_weekly)
dec_ts_sick_weekly <- decompose(ts_sick_weekly)

# Plot the decompositions
# Vacation
plot(dec_ts_vacation_weekly)
```




```r
# Sickness
plot(dec_ts_sick_weekly)
```



Observations:

Looking at the trend components for both vacation and sickness we notice:

- Vacation decreases significantly in 2014 and continues with a smaller trend in 2015, picking up again in 2016
- Sickness shows a slight initial trend decrease, followed by an increase over the years.

Let's explore now the data for both Exception and Productive Hours

Exception vs. Productive Hours

Considering first Providence Health Care as a whole, not making any distinction among facilities, program, and job families, for example.

Analyze the exceptions occurred from 2013 to 2017, contrasting them with the productive hours in order to see if there is a correlation between them.

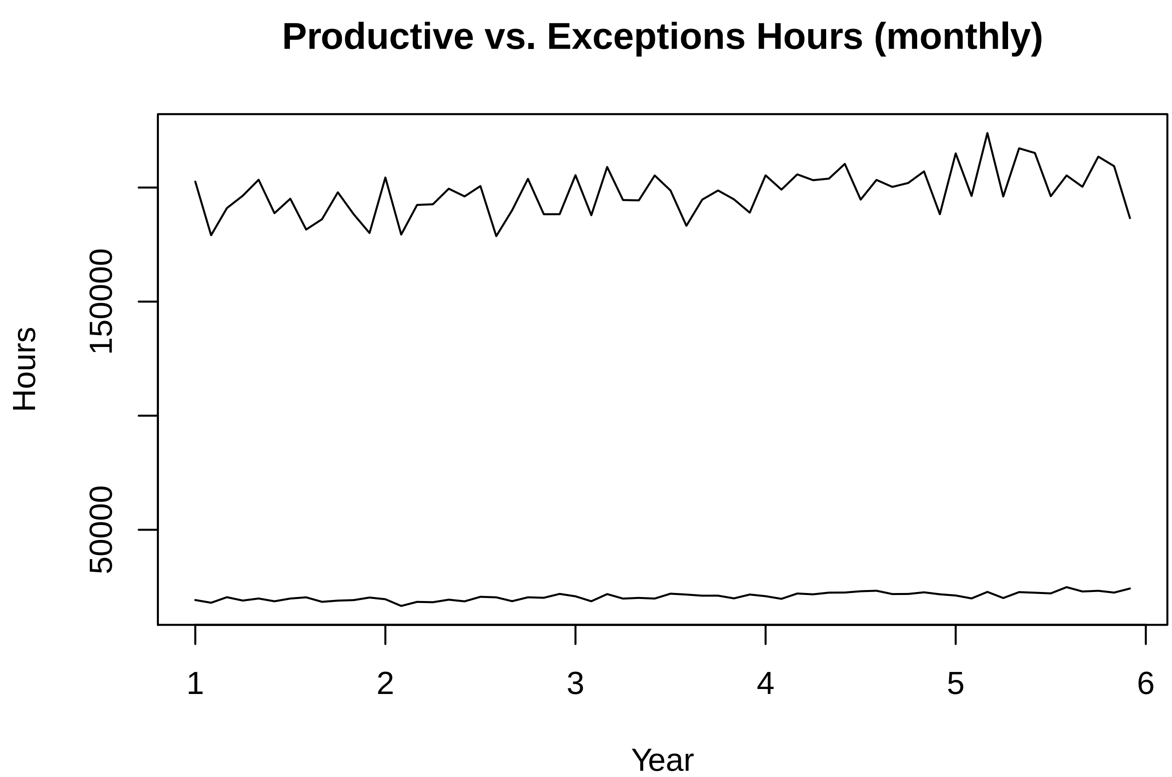
- **Monthly Analysis**

```r
# Create a dataset considering a monthly basis
excep_prod_hours_monthly <- exception_and_productive_hours %>%
  # Consider the same window of the training set - data from 2013 to 2017
  filter(year(SHIFT_DATE) < 2018) %>%
  # extract year and month
  mutate(year = year(SHIFT_DATE),
         month = month(SHIFT_DATE)) %>%
  group_by(year, month) %>%
  summarise(prod_hours = sum(WORKED_HRS),
```

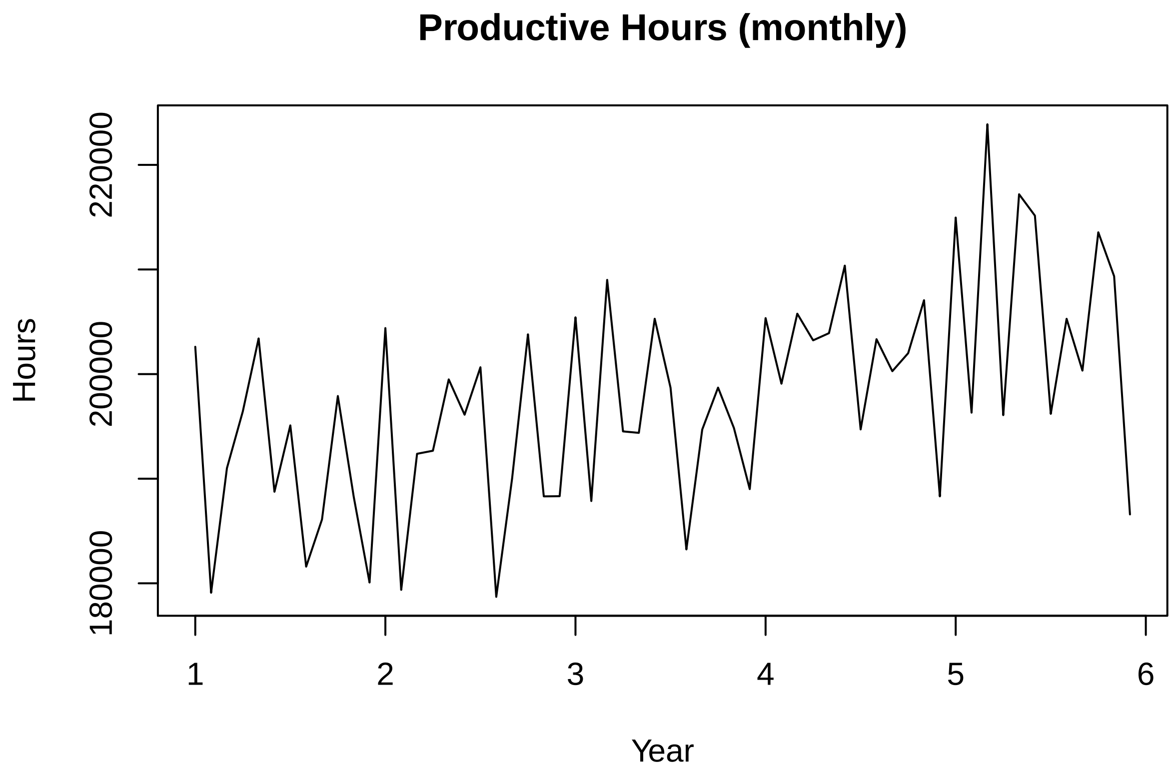
```r
            excep_hours = sum(total_exception_hours),
            total_exceptions = sum(number_of_exceptions))
```

```r
# Create monthly time series
ts_prod_hours_monthly <- ts(excep_prod_hours_monthly$prod_hours, frequency = 12)
ts_excep_hours_monthly <- ts(excep_prod_hours_monthly$excep_hours, frequency = 12)
ts_excep_number_monthly <- ts(excep_prod_hours_monthly$total_exceptions, frequency = 12)
```

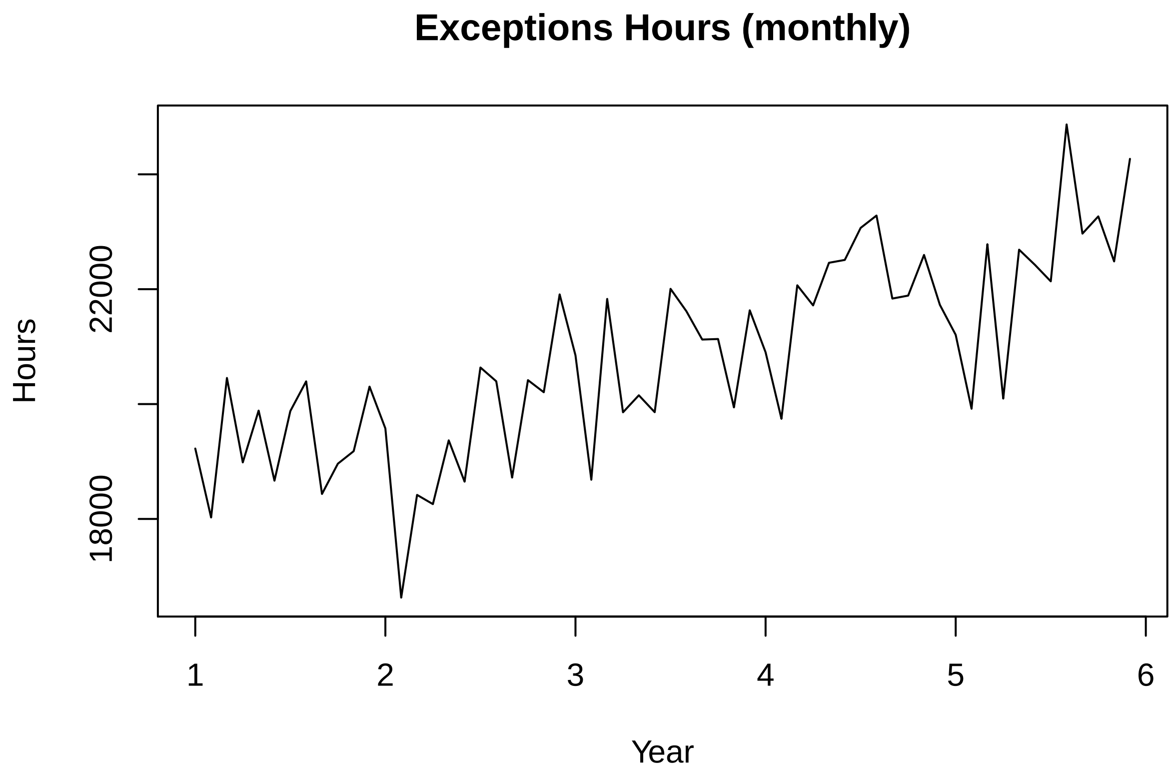
```r
# Plot the time series
ts.plot(ts_prod_hours_monthly, ts_excep_hours_monthly,
        main = "Productive vs. Exceptions Hours (monthly)", xlab = "Year", ylab = "Hours")
```



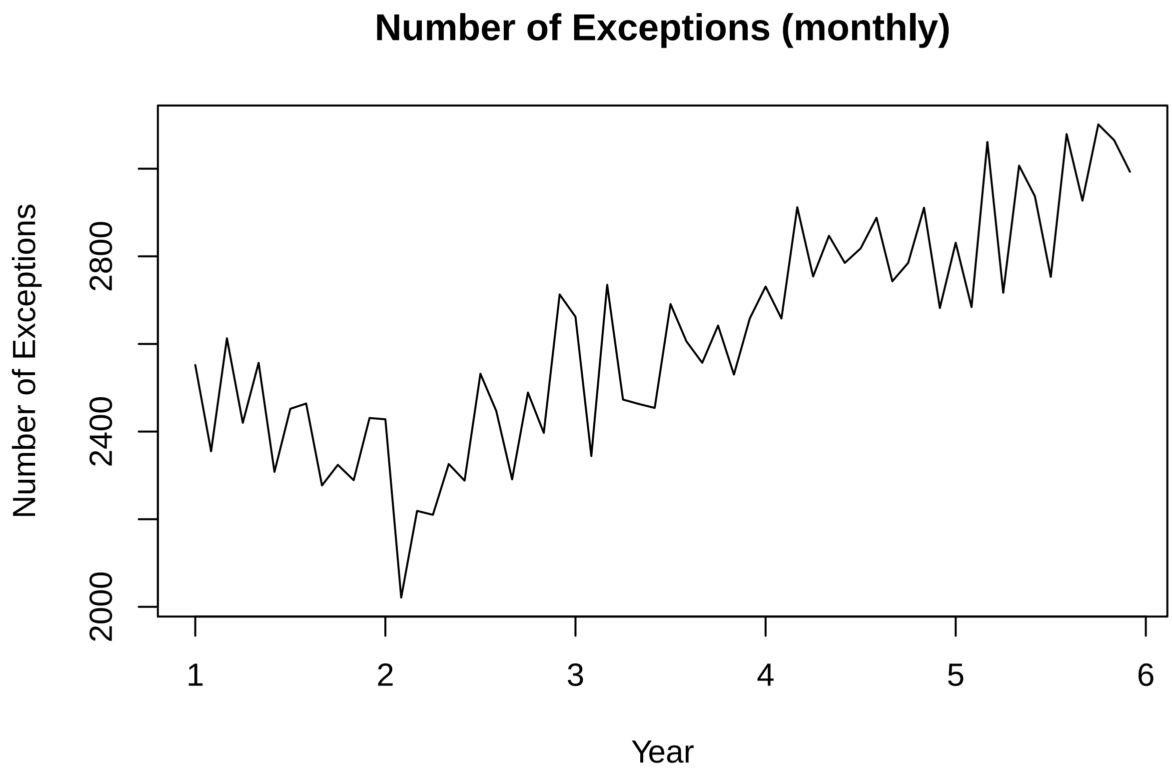

```r
plot(ts_prod_hours_monthly, main = "Productive Hours (monthly)", xlab = "Year", ylab = "Hours")
```



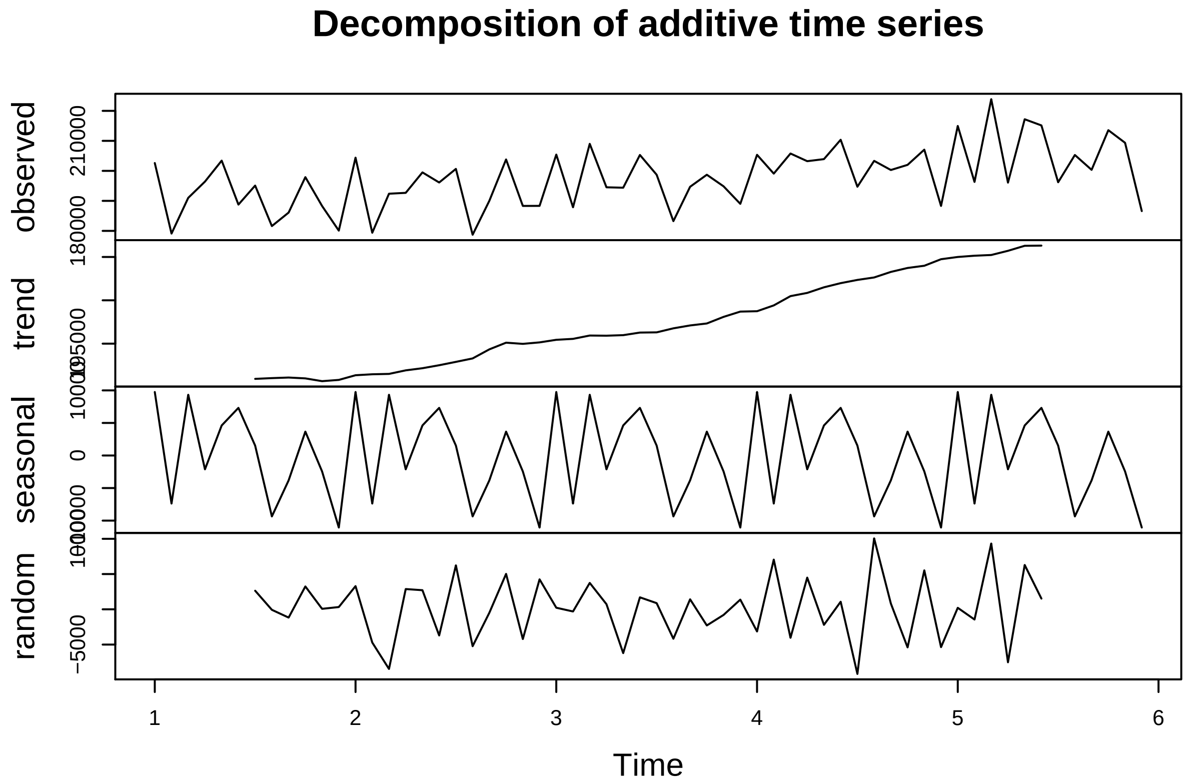

```r
plot(ts_excep_hours_monthly, main = "Exceptions Hours (monthly)", xlab = "Year", ylab = "Hours")
```



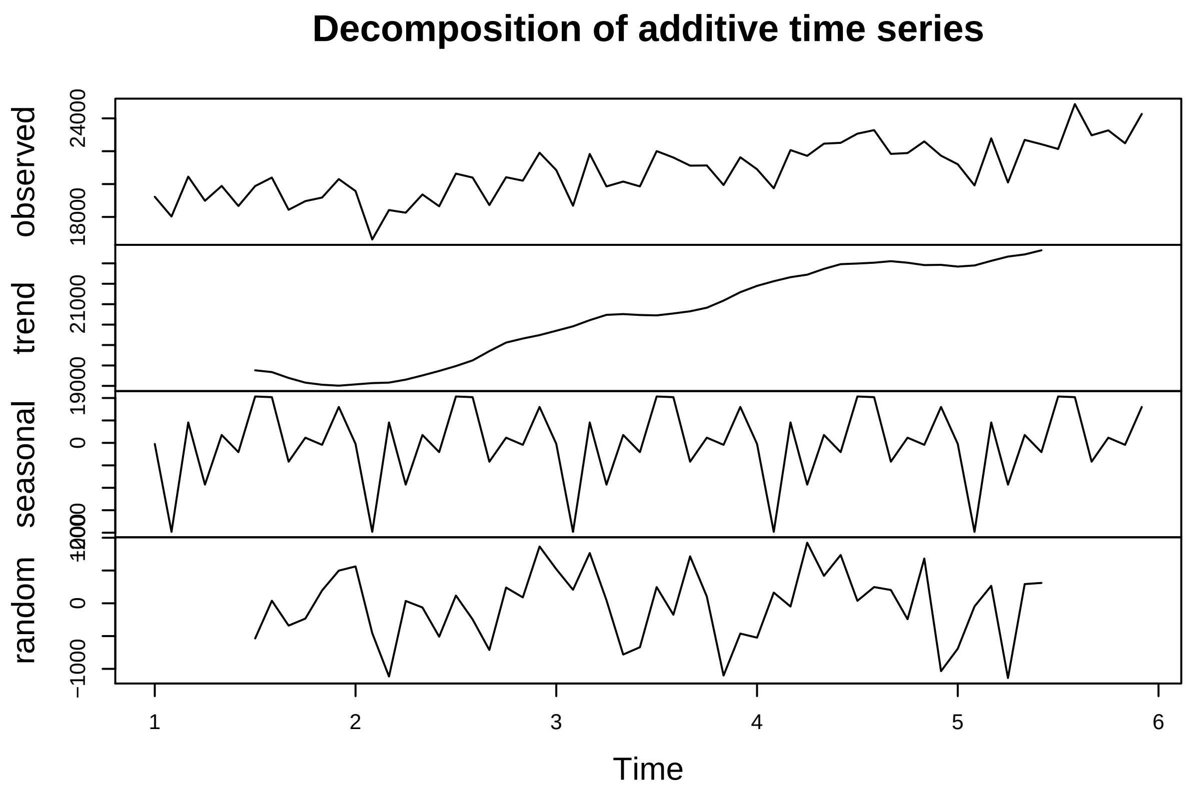

```r
plot(ts_excep_number_monthly, main = "Number of Exceptions (monthly)",
     xlab = "Year", ylab = "Number of Exceptions")
```



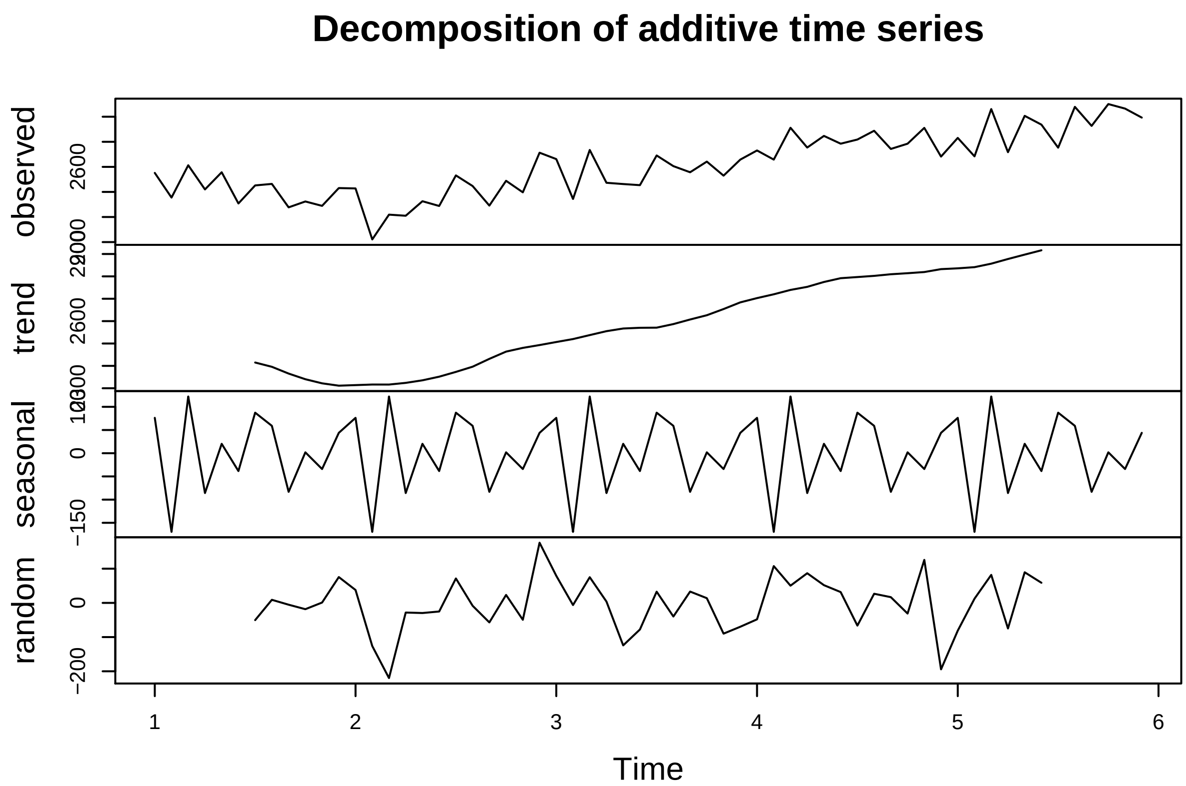

```r
# Plot the decompositions
# Productive hours
dec_ts_prod_hours_monthly <- decompose(ts_prod_hours_monthly)
plot(dec_ts_prod_hours_monthly)
```




```r
# Exception hours
dec_ts_excep_hours_monthly <- decompose(ts_excep_hours_monthly)
plot(dec_ts_excep_hours_monthly)
```




```r
# Total number of exceptions
dec_ts_excep_number_monthly <- decompose(ts_excep_number_monthly)
plot(dec_ts_excep_number_monthly)
```

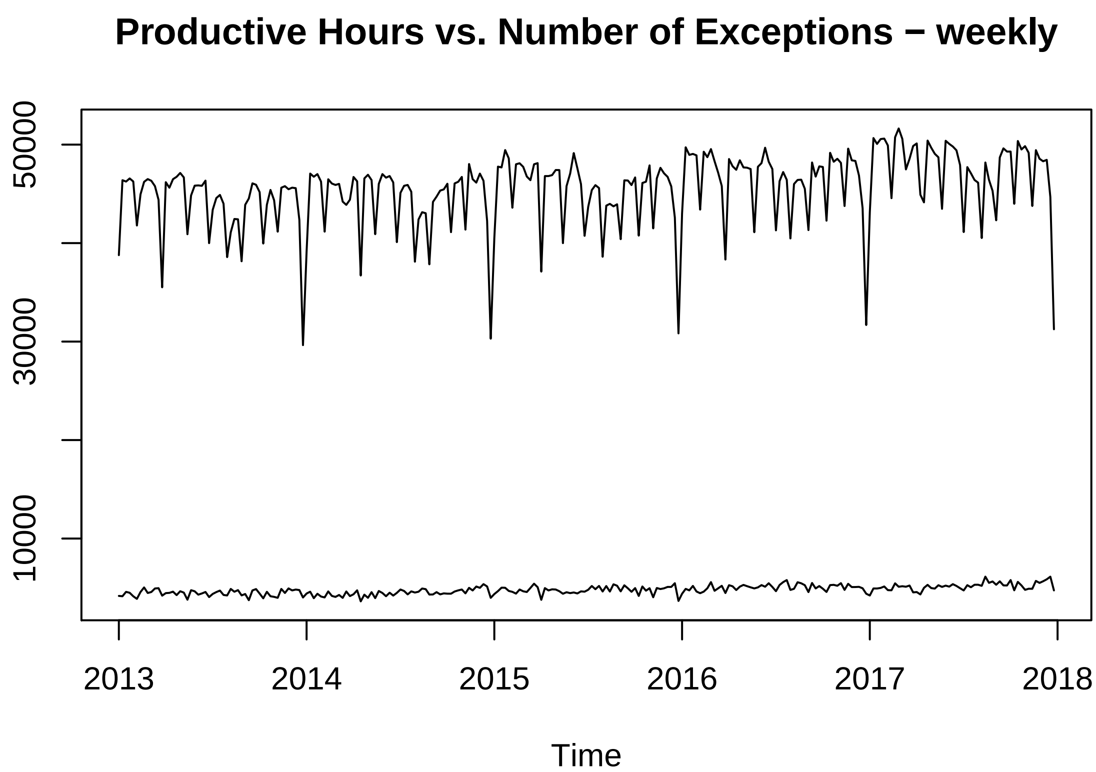


- **Weekly Analysis**

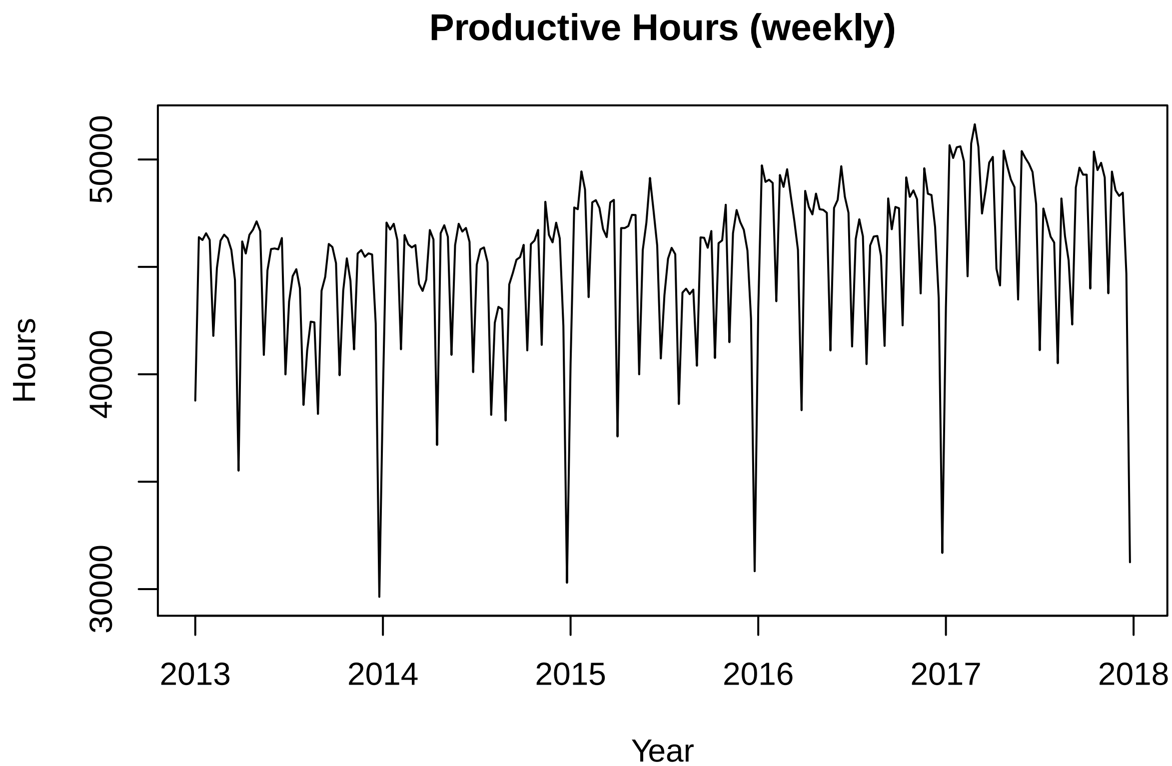
```
# Create a data set considering a weekly basis
excep_prod_hours_weekly <- exception_and_productive_hours %>%
  # Consider the same window of the training set - data from 2013 to 2017
  filter(year(SHIFT_DATE) < 2018) %>%
  # extract year and week
  mutate(year = year(SHIFT_DATE),
         week = week(SHIFT_DATE)) %>%
  group_by(year, week) %>%
  summarise(prod_hours = sum(WORKED_HRS),
            excep_hours = sum(total_exception_hours),
            total_exceptions = sum(number_of_exceptions)) %>%
  # remove the last week of each year (week 53), since they consider few days
  filter(week != 53)
```

```
# Create weekly time series
ts_prod_hours_weekly <- ts(excep_prod_hours_weekly$prod_hours,
                           start = c(2013, 1),
                           frequency = 52)
ts_excep_hours_weekly <- ts(excep_prod_hours_weekly$excep_hours,
                            start = c(2013, 1),
                            frequency = 52)
ts_excep_number_weekly <- ts(excep_prod_hours_weekly$total_exceptions,
                             start = c(2013, 1),
                             frequency = 52)
```

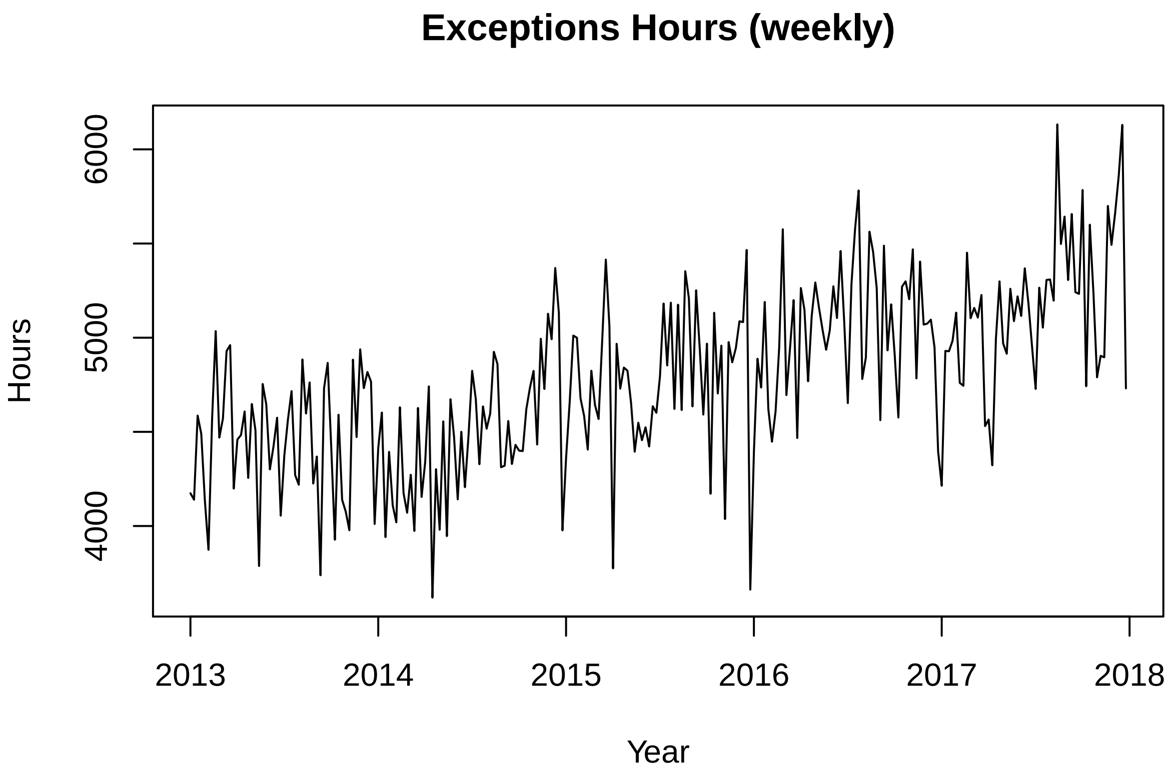
```r
# Plot the time series
ts.plot(ts_prod_hours_weekly,
        ts_excep_hours_weekly,
        main = "Productive Hours vs. Number of Exceptions - weekly")
```



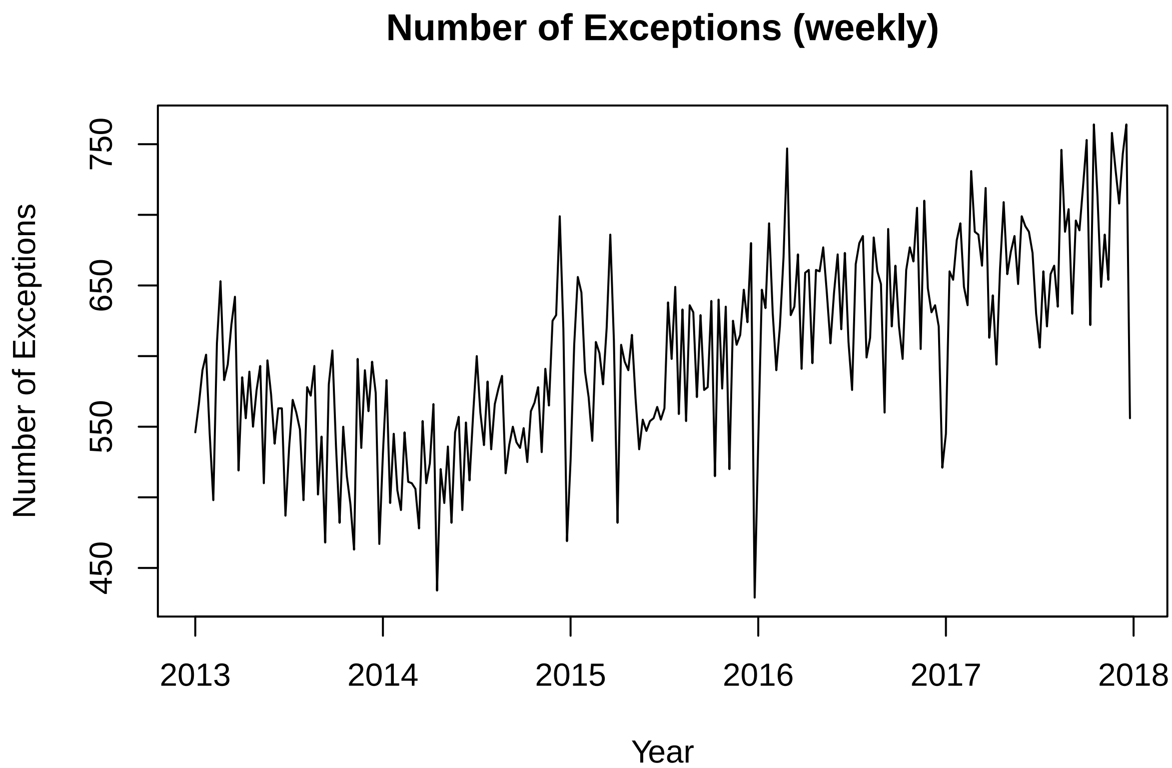

```r
plot(ts_prod_hours_weekly,
     main = "Productive Hours (weekly)",
     xlab = "Year",
     ylab = "Hours")
```



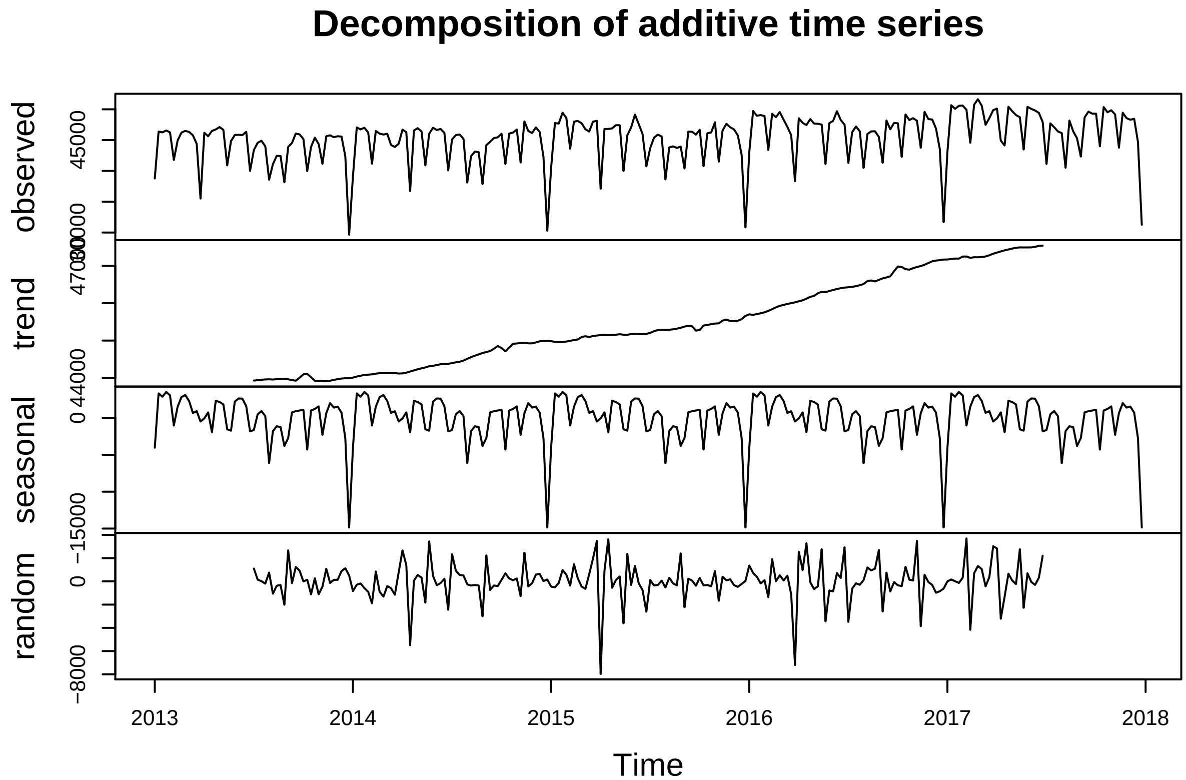

```r
plot(ts_excep_hours_weekly,
     main = "Exceptions Hours (weekly)",
     xlab = "Year",
     ylab = "Hours")
```



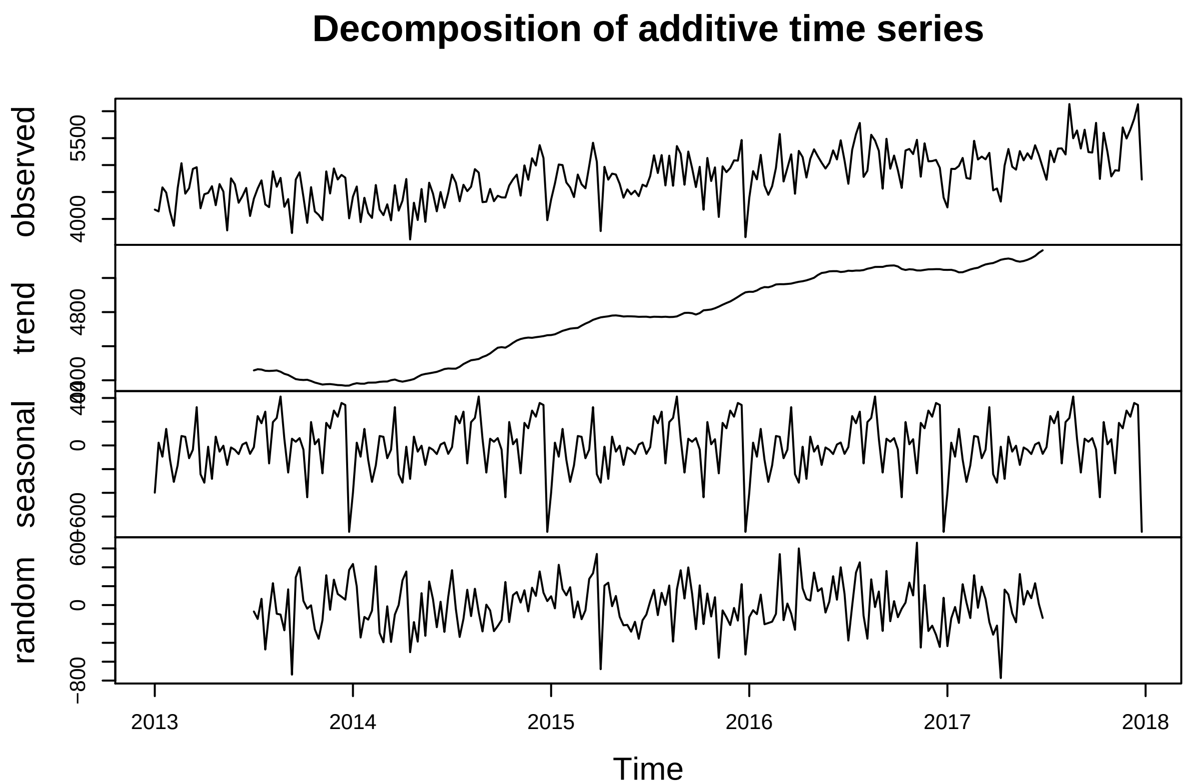

```r
plot(ts_excep_number_weekly,
     main = "Number of Exceptions (weekly)",
     xlab = "Year",
     ylab = "Number of Exceptions")
```



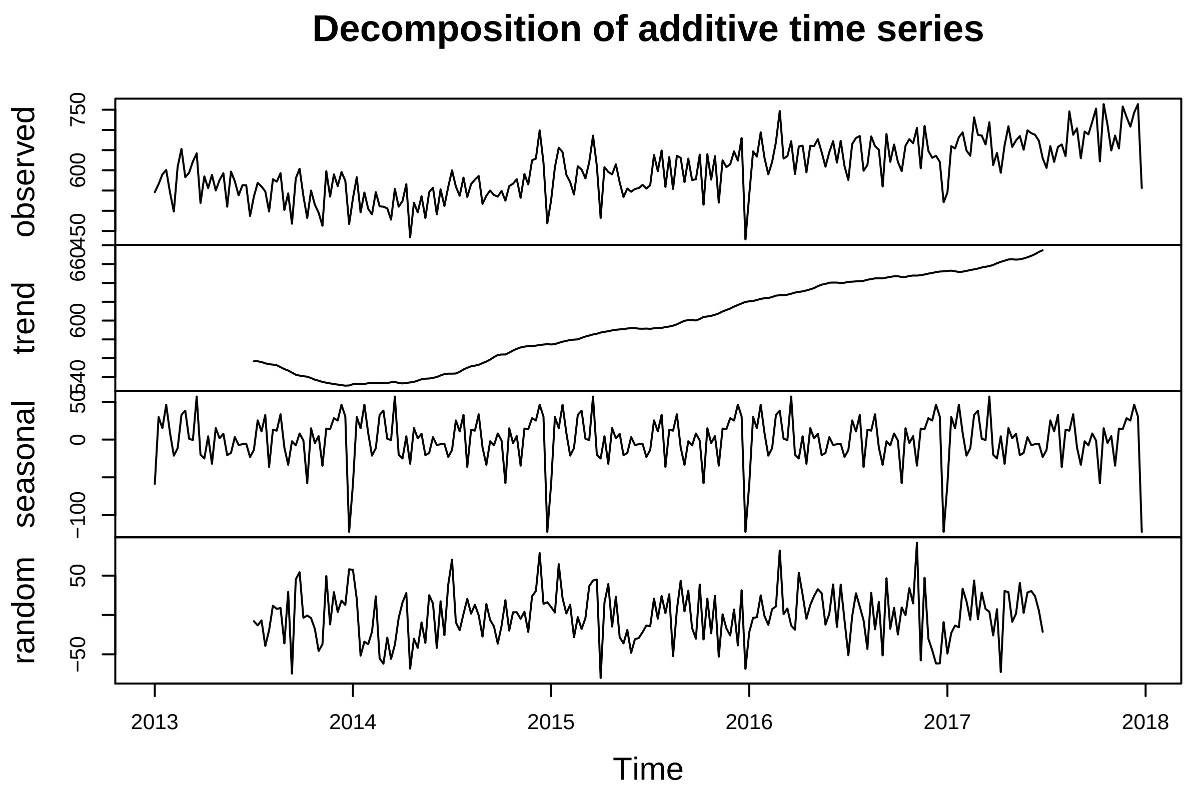

```r
# Plot the decompositions
# Productive hours
dec_ts_prod_hours_weekly <- decompose(ts_prod_hours_weekly)
plot(dec_ts_prod_hours_weekly)
```




```
# Exception hours
dec_ts_excep_hours_weekly <- decompose(ts_excep_hours_weekly)
plot(dec_ts_excep_hours_weekly)
```




```r
# Total number of exceptions
dec_ts_excep_number_weekly <- decompose(ts_excep_number_weekly)
plot(dec_ts_excep_number_weekly)
```



Observations:

- All analyses (monthly and weekly for productive hours, exception hours and number of exceptions) indicate an increasing trend over the years.
- The weekly analyses show that the seasonal component has a trough every year during week 52, i.e. much lower numbers for productive hours, exception hours and number of exceptions in comparison to other surrounding weeks.

```
excep_prod_hours_weekly %>%
  filter(week %in% c(50, 51, 52, 1, 2), !(week %in% c(1, 2) & year == 2013))
```

```
## # A tibble: 23 x 5
## # Groups:   year [5]
##     year  week prod_hours excep_hours total_exceptions
##    <dbl> <dbl>      <dbl>       <dbl>            <dbl>
##  1  2013    50     45577.       4818.              596
##  2  2013    51     42356.       4766.              574
##  3  2013    52     29639.       4011.              467
##  4  2014     1     39126.       4414.              531
##  5  2014     2     47063.       4602.              583
##  6  2014    50     46329.       5370.              699
##  7  2014    51     42251.       5131.              619
##  8  2014    52     30295.       3977.              469
##  9  2015     1     40492.       4360.              526
## 10  2015     2     47767.       4658.              608
## # ... with 13 more rows
```

- At a glance, the expectation would be for weeks with lower productive hours to have higher exceptions (number and/or hours). However, this doesn't seem to be true for week 52, as all values are lower than other weeks.

Do holidays play a part in this? That is, do weeks that have holidays have lower productive hours, but also lower exceptions? Are holidays not taken into account in exceptions?