

# Test 3.1 Bike\_numerical\_&\_graphical\_analysis

Jimmy Janssen van Raay

9-12-2020

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

```
setwd("~/Documenten/Assignment/Bike/R code")
```

### Packages insert (code hidden)

### Inlezen data (code hidden)

### Data Summary

```
#glimpse(BikeData[1:7,])
# Functie 'Glimpse' geeft snel overzicht van de verschillende variabelen en haar data typen
# in de data-set.
```

## Beschrijving variabelen:

- Kolom 1 geeft de duration per gehuurde fiets in seconden.
- Kolom 2 en 3 de begindatum en einddatum van de huur (incl. tijd; kolom 15 & 16 hiervan afgeleid)
- Kolom 4 en 6 geven de station nummer weer, 5 & 7 de naam/adres van het desbetreffende verhuurstation.
- Kolom 8 geeft weer of het om een 'member' gaat of dat er 'casual', dus zonder lidmaatschap wordt gehuurd.
- Vanaf de 9e kolom zijn het aangemaakte variabelen, gedestileerd uit de bestaande, oorspronkelijke data (data engineered) om een vollediger analyse te kunnen maken. Zie hiervoor deel 2 van de R-code van dit project.
- Kolom 9 en 10 zijn de dagen van de week waarop verhuurd wordt voor de begin,- en einddag; hierbij is maandag 1, dinsdag 2 etc.
- Kolom 11, 12, 13, 14 en 21 zijn zogenaamde dummy of indicator variabelen (0 of 1), ook wel bekend als one-hot coding.
- Kolom 11 & 12 geven aan of het al dan niet om een weekenddag gaat(1) of niet (0), voor begin en einddag van de verhuur.
- Kolom 13 geeft aan of de huurder een member is (1) of niet (0).
- Kolom 14 geeft aan of de fiets dezelfde dag wordt teruggebracht (1) of niet (0).
- Kolom 15 & 16 zijn dus de tijden waarop de huur begint en eindigt (afgeleid van kolom 2 & 3).

- kolom 17 & 18 zijn de dagdelen (ochtend=1, middag=2, avond=3 en nacht is 4) waarop begin & eind van de verhuur geschiedt.
- kolom 19 is de duration van de huuris een leesbaarder formaat.
- Kolom 20 en 22 zijn de maand respectievelijk de week (weeknummer) waarin de verhuur geschied.
- Kolom 21 geeft aan of de fiets op hetzelfde station wordt teruggebracht (1) of niet (0).

```
#summary(BikeData)
#summary(BikeData[, c(1:3, 13, 14, 17, 18, 21)])
```

Inzake Summary-function, eigenlijk is van de statistische samenvatting alleen de duration (lengte van rit) interessant en informatief, hieronder zal wel het aantal ritten per tijdseenheid (dag(deel)/week/maand/kwartaal/jaar) worden geanalyseerd, de afstanden en locaties (sectors) worden in # 3.2 geanalyseerd bij de geo-analyse.

De minimale huur per fiets is dus 1 minuut, mediaan zit rond de 600 minuten (10 minuten) en het gemiddelde zit hier stuk boven met 1054 seconden (kwartier). Het betreft dus voornamelijk korte stadsritjes, waarschijnlijk meestens forenzen-ritjes. De maximale verhuur is nog net geen etmaal.

```
skim(BikeData)
```

Table 1: Data summary

|                        |          |
|------------------------|----------|
| Name                   | BikeData |
| Number of rows         | 400000   |
| Number of columns      | 22       |
| <hr/>                  |          |
| Column type frequency: |          |
| character              | 4        |
| difftime               | 2        |
| numeric                | 14       |
| POSIXct                | 2        |
| <hr/>                  |          |
| Group variables        | None     |

#### Variable type: character

| skim_variable   | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|-----------------|-----------|---------------|-----|-----|-------|----------|------------|
| Start_station_5 | 0         | 1             | 10  | 64  | 0     | 480      | 0          |
| End_station_7   | 0         | 1             | 10  | 64  | 0     | 480      | 0          |
| Member_type_8   | 0         | 1             | 6   | 7   | 0     | 3        | 0          |
| RentTime_19     | 0         | 1             | 5   | 11  | 0     | 10396    | 0          |

#### Variable type: difftime

| skim_variable | n_missing | complete_rate | min    | max        | median     | n_unique |
|---------------|-----------|---------------|--------|------------|------------|----------|
| Start_time_15 | 0         | 1             | 0 secs | 86399 secs | 54790 secs | 70526    |
| End_time_16   | 0         | 1             | 0 secs | 86399 secs | 55957 secs | 70695    |

#### Variable type: numeric

| skim_variable           | n_missing | complete_rate | mean    | sd      | p0       | p25     | p50      | p75     | p100  | hist |
|-------------------------|-----------|---------------|---------|---------|----------|---------|----------|---------|-------|------|
| Duration_1              | 0         | 1             | 1054.37 | 1984.72 | 60       | 388.00  | 656.0    | 1113.00 | 86223 |      |
| Start_station_number_04 | 1         | 31304.84      | 205.67  | 31000   | 31202.00 | 31243.0 | 31404.00 | 32225   |       |      |
| End_station_number_06   | 1         | 31306.01      | 203.40  | 31000   | 31203.00 | 31243.0 | 31402.00 | 32225   |       |      |
| Weekday_Start_9         | 0         | 1             | 3.97    | 1.97    | 1        | 2.00    | 4.0      | 6.00    | 7     |      |
| Weekday_End_10          | 0         | 1             | 3.97    | 1.97    | 1        | 2.00    | 4.0      | 6.00    | 7     |      |
| Weekend_Start_11        | 0         | 1             | 0.27    | 0.44    | 0        | 0.00    | 0.0      | 1.00    | 1     |      |
| Weekend_End_12          | 0         | 1             | 0.27    | 0.44    | 0        | 0.00    | 0.0      | 1.00    | 1     |      |
| Member_13               | 0         | 1             | 0.79    | 0.40    | 0        | 1.00    | 1.0      | 1.00    | 1     |      |
| Samedayback_14          | 0         | 1             | 1.00    | 0.06    | 0        | 1.00    | 1.0      | 1.00    | 1     |      |
| Day_part_Start_17       | 0         | 1             | 1.78    | 0.75    | 0        | 1.00    | 2.0      | 2.00    | 3     |      |
| Day_part_End_18         | 0         | 1             | 1.82    | 0.77    | 0        | 1.00    | 2.0      | 2.00    | 3     |      |
| Month_20                | 0         | 1             | 6.50    | 3.33    | 1        | 3.75    | 6.5      | 9.25    | 12    |      |
| Zelfde_station_21       | 0         | 1             | 0.03    | 0.18    | 0        | 0.00    | 0.0      | 0.00    | 1     |      |
| WeekNr_22               | 0         | 1             | 26.01   | 14.47   | 0        | 13.00   | 26.0     | 39.00   | 53    |      |

### Variable type: POSIXct

| skim_variable | n_missing | complete_rate | min                 | max                 | median              | n_unique |
|---------------|-----------|---------------|---------------------|---------------------|---------------------|----------|
| Start_date_2  | 0         | 1             | 2012-01-01 00:50:36 | 2017-12-31 20:22:46 | 2015-06-15 17:51:22 | 399091   |
| End_date_3    | 0         | 1             | 2012-01-01 00:53:00 | 2017-12-31 20:28:05 | 2015-06-15 18:06:33 | 399086   |

Met de functie 'skim' krijg je een aardig overzicht van de data inclusief eenvoudige plots, handig voor eerste indruk. In eerste instantie zijn alle bestanden (kwartalen en jaren) met skim bekijken maar beter is om deze af te zetten in een grafiek met datgene (bv duration) met wat je nader wilt analyseren. De grafieken vallen helaas wel weg in de pdf.

## Plot Summaries

### Duration-plots general

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 318 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

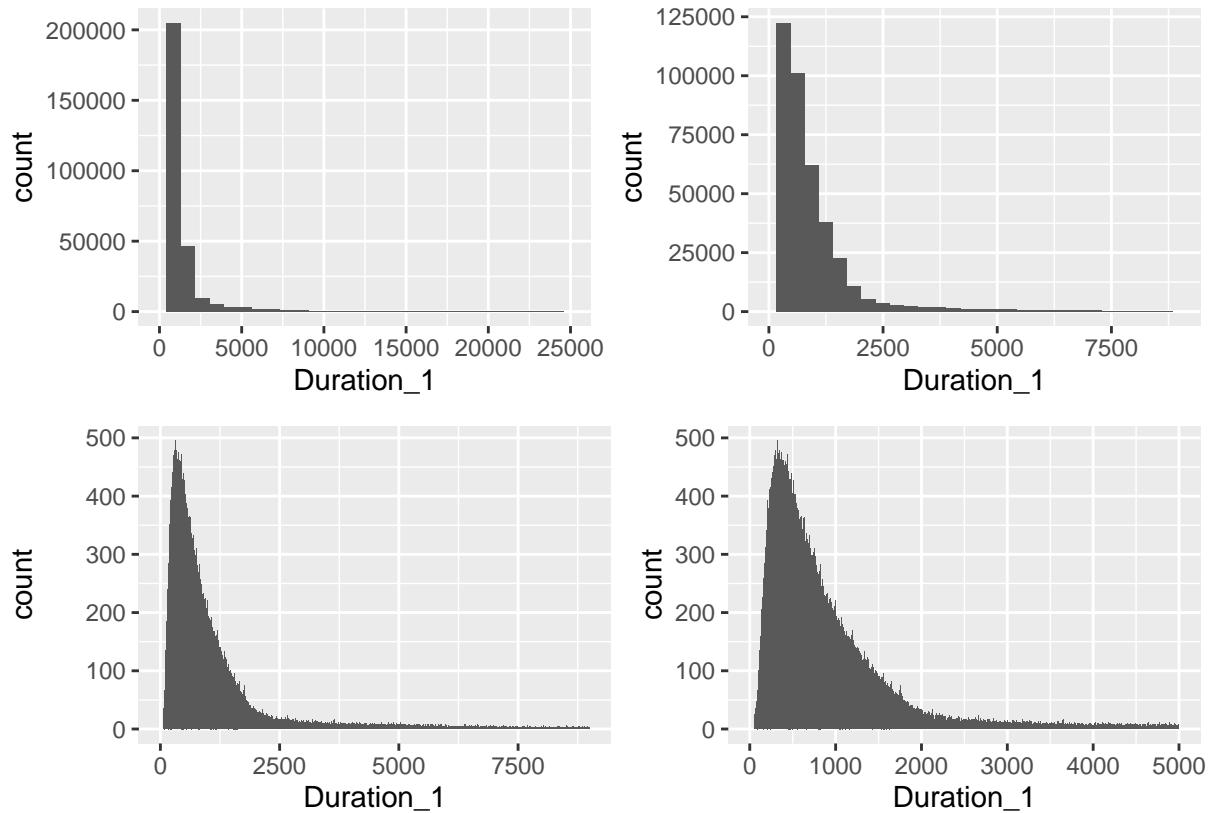
## Warning: Removed 2690 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).

## Warning: Removed 2690 rows containing non-finite values (stat_count).

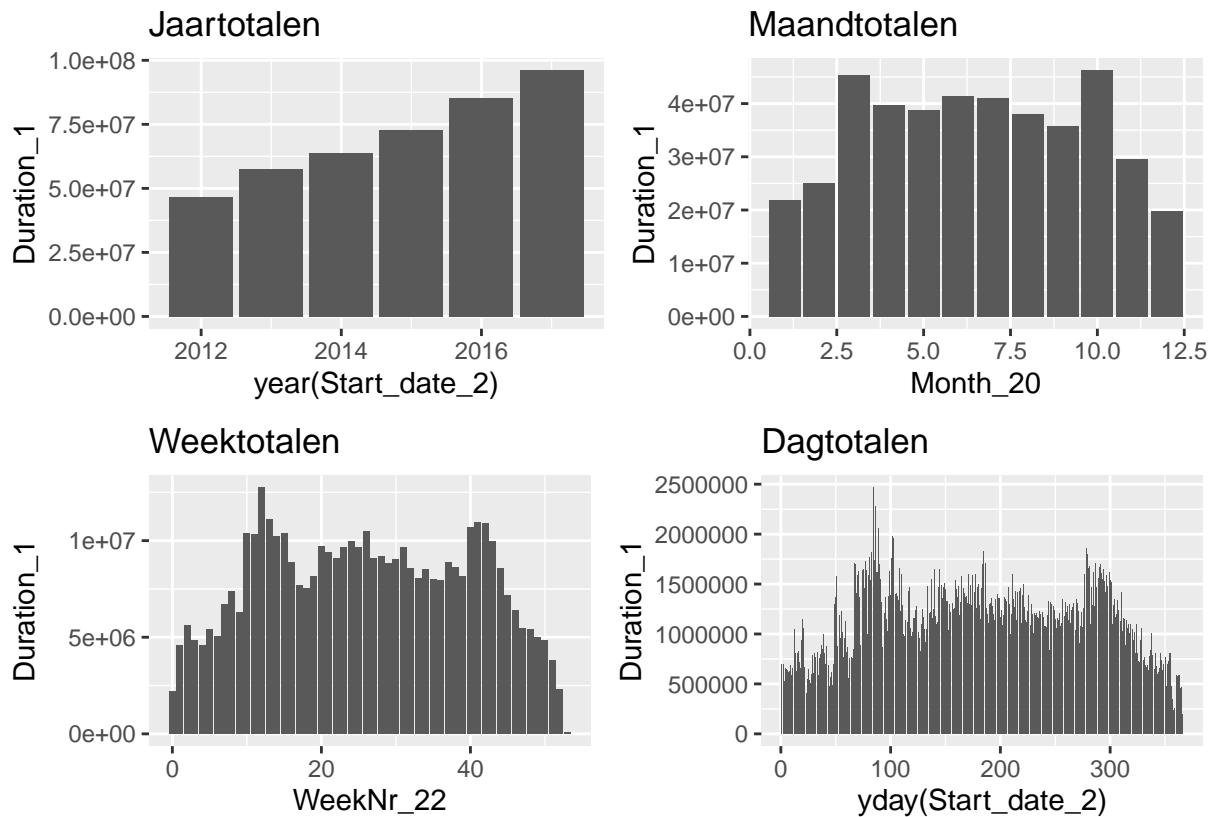
## Warning: Removed 9839 rows containing non-finite values (stat_count).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



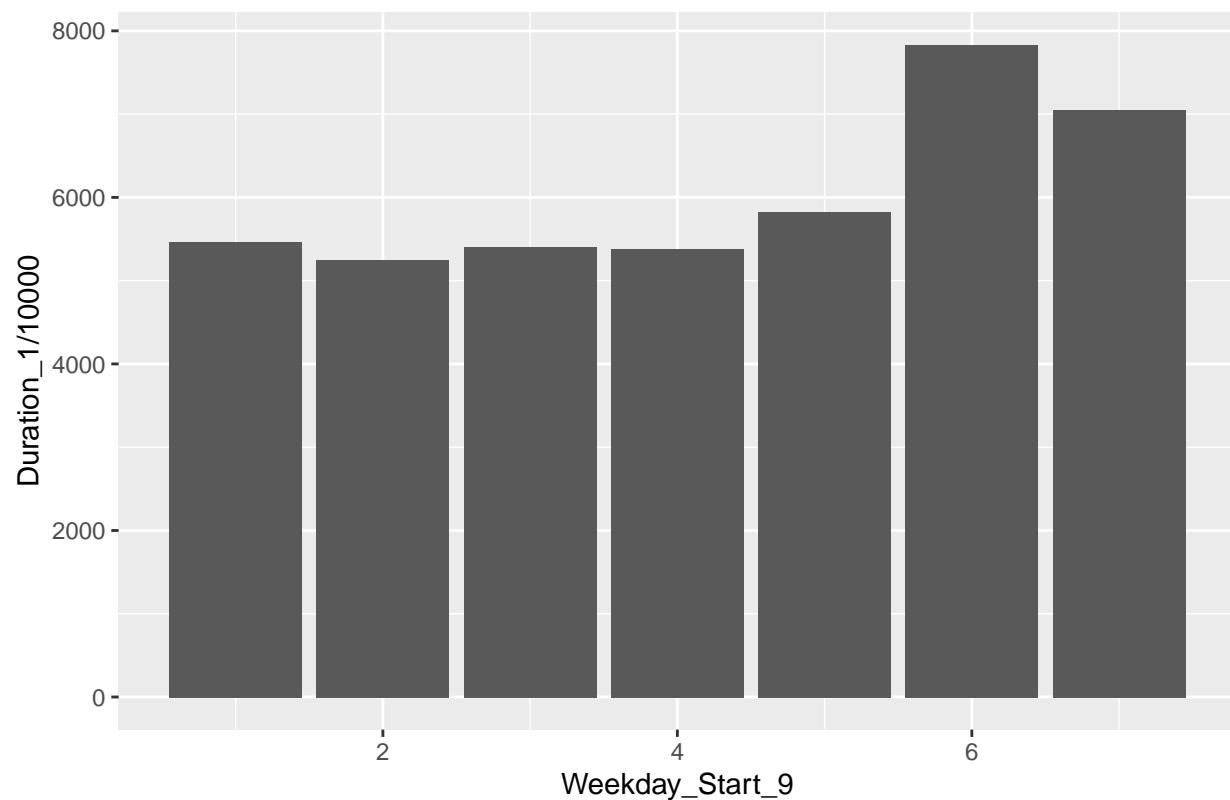
Het grootste deel van de verhuur vindt binnen hetzelfde etmaal ( $<1440$ ) plaats, gemiddeld ( $n=920$ ) genomen duren de meeste trips wel langer dan een halve dag ( $>720$ ) maar qua aantal/mediaan zijn de meeste trips zelfs korter ( $m=587$ ). Wel zijn er enige enorme uitschieters naar boven toe.

### Duration-plots per selected variable



- Bij het afzetten van de het jaartal t.o.v. de duration is mooi de groei van jaar op jaar te zien.
- Bij de grafiek van maand t.o.v. duration is ook in zekere zin een ‘logische’ opbouw te zien richting zomer met wel met maart & oktober als drukste maand (waarschijnlijk minste vakantie tijdens deze maanden).
- Bij de grafiek met duration-duur per week is nog duidelijker de piek in het voorjaar en het najaar te zien.
- Rond de jaarwisseling is verreweg de minste duration/verhuur.
- De dag-grafiek laat een nog mooier grilliger beeld zien. Hierin zitten ongewijfeld feestdagen en zal ook het weer van de dag een rol hebben gespeeld!

Weekverloop start



Weekverloop eind

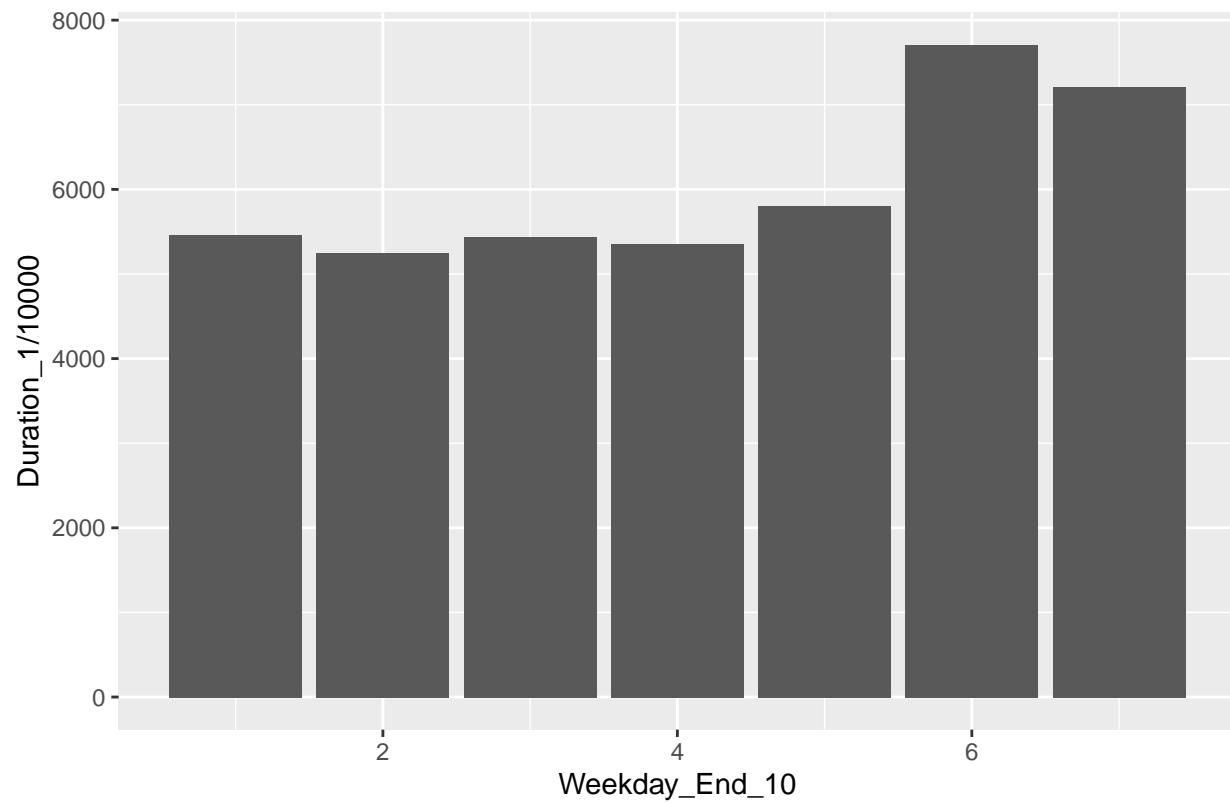


Table 6: Overzicht huurders lidmaatschap, totaal en percentage

| Lid    | Freq    | Lid    | Freq     |
|--------|---------|--------|----------|
| Casual | 82,193  | Casual | 0.205483 |
| Member | 317,806 | Member | 0.794517 |

Table 7: Overzicht huurders per weekdag, aantal en prop

|        | 1      | 2      | 3      | 4      | 5      | 6      | 7      |
|--------|--------|--------|--------|--------|--------|--------|--------|
| Casual | 9,579  | 7,818  | 8,024  | 8,208  | 10,679 | 19,906 | 17,979 |
| Member | 45,726 | 50,024 | 51,828 | 50,963 | 48,468 | 36,169 | 34,628 |
|        | 1      | 2      | 3      | 4      | 5      | 6      | 7      |
| Casual | 0.024  | 0.020  | 0.02   | 0.021  | 0.027  | 0.05   | 0.045  |
| Member | 0.114  | 0.125  | 0.13   | 0.127  | 0.121  | 0.09   | 0.087  |

- De laatste 2 grafieken met duration per dagdeel laten een duidelijk drukkere verhuur zien in het weekend.
- Blijkbaar zijn het niet alleen forenzen die fietsen huren!
- Verschil binnen dag tussen begin en einde verhuur is (zoals te verwachten valt, klein.)

#### Duration-tables & proportion-tables per selected variable

### Member tables

- Bijna 80% van alle huurders betreft leden van het verhuurfietsplan.

|        | ma    | di    | wo    | do    | vrij  | za    | zo    |
|--------|-------|-------|-------|-------|-------|-------|-------|
| Casual | 17.36 | 13.56 | 13.39 | 13.85 | 17.99 | 35.44 | 34.24 |
| Member | 82.64 | 86.44 | 86.61 | 86.15 | 82.01 | 64.56 | 65.76 |

- In het weekend is de verhouding tussen incidentele gebruikers en leden meer in balans, ruwweg 35% vs. 65%, in plaats van 17% vs. 83%
- Enkel de verhouding werkdag en weekenddag geeft een nog iets beter beeld tussen weekend en doordeweekse drukte en vooral de verhouding.
- Grootste deel verhuur is leden doordeweeks (61%), gevolgd door leden weekend (17,7%) en daarna niet-leden met zo'n 10% verdeeld over week,- en weekend.
- Bijna alle fietsen worden dezelfde dag teruggebracht, dit viel te verwachten gezien de aard van de fietsverhuur (grab&go).
- Verhuur opgesplitst naar dagdelen geeft een goed beeld wanneer de drukste dagdelen zijn.
- Nacht is veruit het rustigste moment, ochten en vooral de middag zijn de drukste dagdelen.

| SSt | 7     | Freq |
|-----|-------|------|
| 0   | 0.966 |      |
| 1   | 0.034 |      |

Table 10: Verhouding zelfde dag terug (1) vs. dag later(0), aantal en prop

|        | 0   | 1       |        | 0     | 1     |
|--------|-----|---------|--------|-------|-------|
| Casual | 663 | 81,530  | Casual | 0.002 | 0.204 |
| Member | 717 | 317,089 | Member | 0.002 | 0.793 |

Table 11: Totalen per dagdeel, begin/eind verhuur

|        | nacht | ocht    | middg   | avond  |        | nacht | ocht    | middg   | avond  |
|--------|-------|---------|---------|--------|--------|-------|---------|---------|--------|
| Casual | 1,560 | 23,827  | 43,604  | 13,202 | Casual | 1,560 | 23,827  | 43,604  | 13,202 |
| Member | 6,314 | 118,436 | 134,726 | 58,330 | Member | 6,314 | 118,436 | 134,726 | 58,330 |

## Time-Graphs Analyses

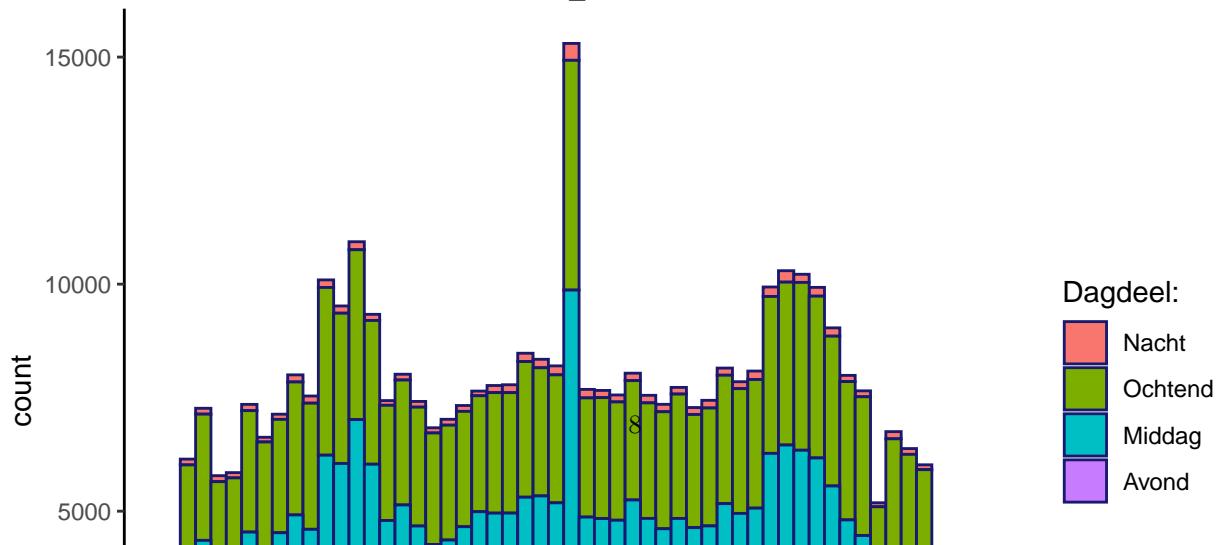
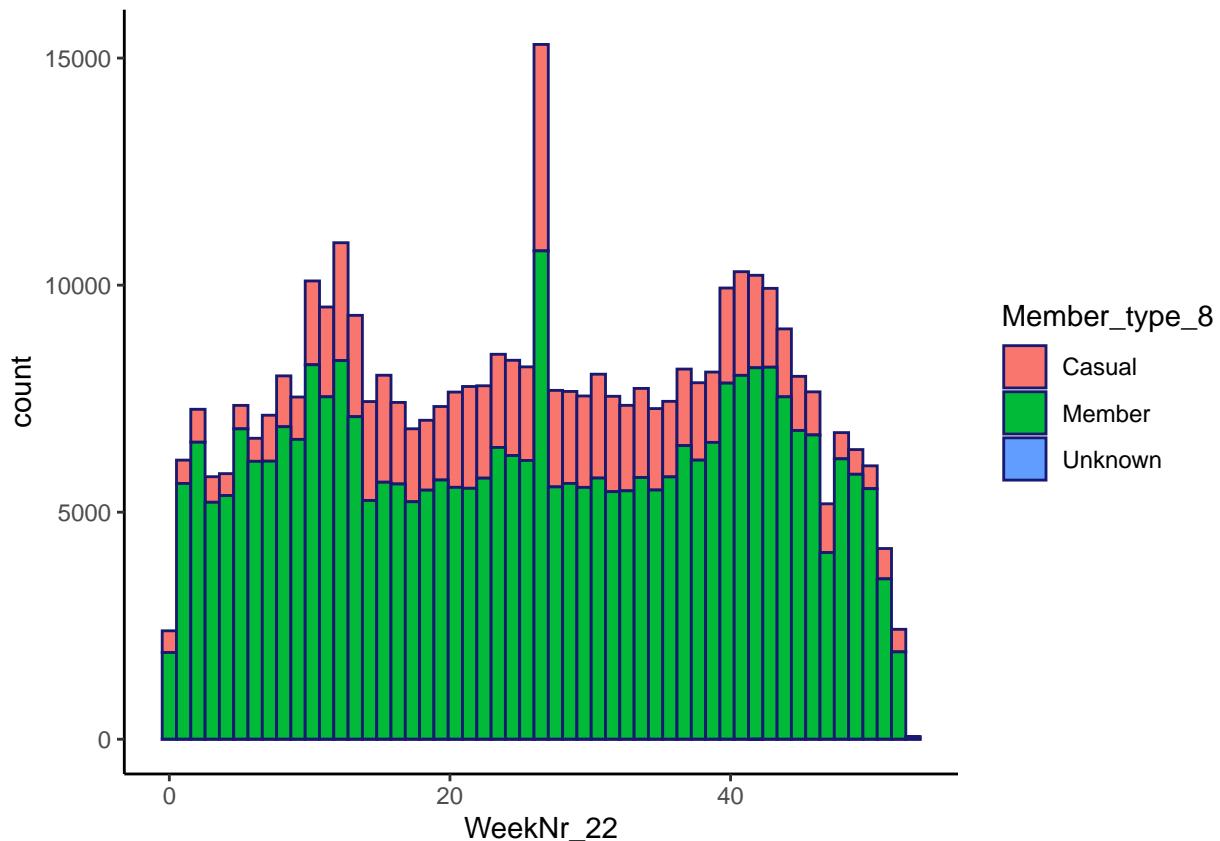
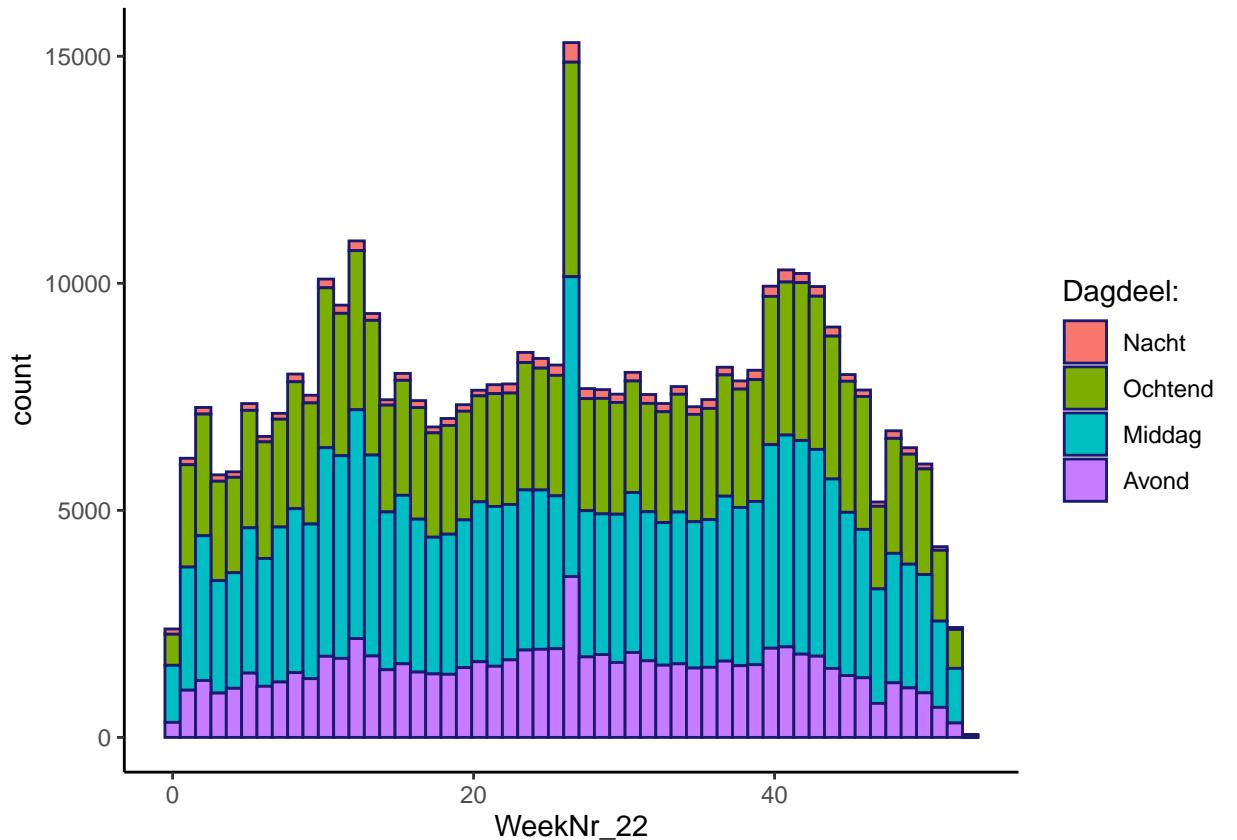


Table 13: Zelfde station terug, ja=1

|        | 0       | 1     |        | 0     | 1     |
|--------|---------|-------|--------|-------|-------|
| Casual | 73,465  | 8,728 | Casual | 0.184 | 0.022 |
| Member | 313,046 | 4,760 | Member | 0.783 | 0.012 |



Op de grafieken in aanvulling op de tabelinformatie is nog eens goed visueel het verschil te zien in member & casual-huurders. \* Tevens is ook visueel inzichtelijk(-er) de verdeling over de 3 dagdelen.