Data Science and Marketing Analytics

Assignment 1:

Dataset

October 2nd 2017

Group 10

Joan Rosabella	**S3296687**
Frans Simanjuntak	**S3038971**
Carlos Humberto Paz R	**S3040577**

In this assignment, we created a dataset that includes relevant Key Performance Indicators (KPIs) of beachwear sales in **wehkamp.nl**. The aim was to prepare a dataset to generate insights in how weather condition drive consumer behavior towards beachwear.

Description of the dataset and motivation

Firstly, we defined the "weather" as daily average temperature, because according to extensive work experience in the clothing industry, the researchers acknowledge that temperature often affects consumers' purchasing behavior (Bahng & Kincade, 2012). Furthermore, previous study by Parsons (2001) found that although there is a significant correlation of maximum temperature with customer shopping pattern, however other weather variables such as humidity and sunshine hours, failed to meet the significant level. In addition, this study was done in New Zealand, where the climate is approximately similar with the Netherlands. Due to budget constrains for obtaining detailed hourly level data from the major weather websites, we decided to use the publicly available temperature offered by the Royal Netherlands Meteorology Institute (http://www.sciamachy-validation.org/climatology/daily_data/selection.cgi) which has information at the day level, therefore our analysis will be performed at this granularity level.

We are interested also in understanding the demographic influence on our analysis, therefore we decided to classify the data by gender and by age. As explained by Markert (2004), there is similar behaviour between groups of people from similar ages, in practice the most common analysis is by dividing the population in groups of 5,10 or 20 years. In our project, since we do not have a significantly large dataset, we are using the 20 years splitting, therefore using the generation theory proposed by Markert (2004) plus given that we have the "birth year" on the customer table, we will use the following age groups by year of birth:

- 1900 - 1945
- 1946 - 1965
- 1966 -1985
- 1986 - 2005
- 2005 - 2017

A possible improvement for choosing the slicing could be by applying a clustering technique to find groups of ages with similar behaviour.

We determined 5 relevant KPIs as mentioned below :

1. IMPRESSION / PAGEVIEWS

By keeping track of daily webpage views, one can identify the efficacy of marketing strategies, whether the campaign successfully drives consumer traffic (DeMers, 2014). A high page views is a good sign that the customer is engaged, and quite often means that they are coming back regularly to the webpage (Bhapkar, 2013). We obtained the data from the *article event table* by counting the type "10" (view) on the article event type. Thus, we can also see if the fluctuation in page views is correlated with certain condition of weather.

The description of the impression/pageview variables is given in the below table:

Variable	Description
article_event_date	Actual date from January 1st to July 31st, 2017
number_views	Amount of articles from wehkamp.nl site being viewed in a day

2. CONVERSION

Build upon the number of pageview we determined earlier, we calculated the conversion rate for our second KPI. This is one of the most important metrics for measuring the profitability of overall marketing efforts (DeMers, 2014), whether the pageview indeed leads customer to actually buy the product. Consequently, managers will be able to identify weak areas in the page design and content thus make changes to create a more effective website. According to Mummalaneni (2005), conversion rate is proportion of website visitors that actually place a purchase order. Therefore in our assignment, we will calculate it as follow :

$$conversion\ rate\ = \frac{number\ of\ articles\ sold}{number\ of\ articles\ viewed}$$

where both numbers obtained from *article event table* (type "10" for view and type "40" for sales). Eventually, we can determine whether the urge of buying of customer (after being encountered by wehkamp page) is correlated with the weather.

It is important to note that since we are analyzing which visits turned into a sale when the customer clicks the "buy" button, we are not taking into account if the customer returns the item eventually. The number of sales means the number of times an article was sold and not the number of orders that were completely delivered and not returned. In the event that the user buys x number of the same article, it will be counted as one article.

The description of the conversion variable is given in the below table:

Variable	Description
article_event_date	Actual date from January 1st to July 31st, 2017
number_sales	The number of articles sold, article event type "10"
number_views	The number of articles viewed, article event type "40"
sales_conversion	Number or sales made by customer after viewing the site

3. SALES

Sales is the most important role in every business. Not only does it give the information about the cash flow, but also the information such as market demand and sales forecast can also be obtained which might influence the business in the future.

For wehkamp.nl, the information about the sales will also be the most influential one. On this occasion, the sales information that we obtained is how much was the total sales and how many items were sold each day. By looking at the total sales, one can notice whether the business profitable or not and also it helps to forecast the sales in the future. Moreover, the total number of items sold per day indicate the sales growth. We can predict how many sales in the following days, months or year. It also provides the information such as on which day or month the sales reached its highest peak or vice versa that will help us to decide when to boost our marketing strategy (e.g spread the adds or give discount). As the main goal of this assignment is to see whether sales is affected by the weather, it is definitely important to include the actual sales data.

In order to obtain the sales information from the wehkamp database, we decided to use the "items" from *order table*, because it has the detail of the delivery and return ("items" means the "delivered" subtracted by "returned"). We also included the sales in euro amount.

The description of the sales variables is given in the below table:

Variable	**Description**
order_date	Actual date from January 1st to July 31st, 2017
total_items	Total sales in unit
total_sales	Total sales in euro

4. MOST SEARCHED ITEMS

In order to maximise the beachwear sales, managers also would not want to overlook the fashion trend by tracking the most search items from time to time. By making sure the availability of trendy items, managers could attract new customers while keeping the existing ones. Furthermore, it gives a better understanding of the terminology that the existing/potential customers are using to find the products that wehkamp is selling, therefore managers could match the website content and marketing terminology (Lazazzera, 2014). We obtained the most searched item from *onsitesearch table* on daily basis.

The description of the most search item variable is given in the below table:

Variable	**Description**
search_term	The string of the most searched keyword
max	The number of searches for this keyword
onsite_search_date	The date of the searched keyword
rn	An auxiliar field for obtaining only the most searched keyword

5. PERCENTAGE OF RETURNED ITEMS

Since we are analyzing the influence of the weather on the behaviour of the customers, an additional interesting metric is to include the percentage of items returned, in order to find if there is a correlation in the number of customers that regret the purchase and the weather.

$$return\ percentage = \frac{number\ of\ orders\ returned}{number\ of\ orders\ delivered}$$

The description of the percentage of returned items variable is given in the below table:

Variable	**Description**
order_date	Actual date from January 1st to July 31st, 2017
Returned	The number of orders returned
Delivered	The number of delivered orders
PercReturned	The percentage of returned items

FINAL DATASET

For the final dataset, we decided to integrate all of the KPIs into a single SQL query, in order to export the table in a csv file for integrating it finally in R with the weather dataset.
In order to avoid extra calculations during analysis we decided to include some extra fields like total number of returned orders, the count of the most searched term, and segregations of sales by demographic groups.
The weather dataset obtained from the Royal Netherlands Meteorology Institute provided measurements for different stations throughout the Netherlands, however the data is not complete for all stations, plus the data is missing for the dates july 27th to july 31st, therefore we decided to use the temperature file provided during the practical laboratories, which comes from the same website but has the data complete for these days.f we had the address information of the customer, we could have done a more specific analysis with the closest station.

The detailed description of the field in the final dataset is given in the below table:

Variable	**Description**
date	The date of the event
number_items_sold	Total number of items sold
number_page_view	Total number of viewed items (impression KPI)
sales_conversion	The percentage of item views that turned into a sale (conversion KPI)

total_orders	Total numbers of orders that turned into a sale (sales KPI)
sales_euro	Total amount of sales (Sales KPI)
returned	Number of returned orders
delivered	Number of delivered orders
percreturned	Percentage of orders returned (Returned Items KPI)
most_searched_term	Most searched term (Most searched term KPI)
total_num_term_searches	Number of searches for the most searched term
femalesales	Number of sold orders for female population
malesales	Number of sold orders for male population
femalesales_1945	Number of sold orders for female customers born between 1900 and 1945
femalesales_1965	Number of sold orders for female customers born between 1946 and 1965
femalesales_1985	Number of sold orders for female customers born between 1966 and 1985
femalesales_2005	Number of sold orders for female customers born between 1986 and 2005
femalesales_2017	Number of sold orders for female customers born between 2005 and 2017
malesales_1945	Number of sold orders for male customers born between 1900 and 1945
malesales_1965	Number of sold orders for male customers born between 1946 and 1965
malesales_1985	Number of sold orders for male customers born between 1966 and 1985
malesales_2005	Number of sold orders for male customers born between 1986 and 2005
malesales_2017	Number of sold orders for male customers born between 2005 and 2017
avg_temperature	Average daily temperature

The final dataset was tested using R library MICE, in order to verify if there is some missing data. The dataset is 100% complete. The code can be found in the appendix section.

Descriptives of all variables

Statistic Summary

The overall statistic summary of our final dataset is depicted in the below figure. This figure describes the minimum and the maximum, the mean and the median, and the first and the third quarter value of each variable excepts for the most searched term. Some insights can be gathered directly by looking at the summary at first glance, for instance: when we compare the summary between the groups of female customers, it can be easily noticed that the female customers who are above 51 never bought products from wehkamp. Extremely, this phenomenon happens on male customers which is in fact, the male customers above 31 never did shopping on the website.

In the comparison to the other variables, the summary of the variable most_searched_term shows different results. The outcome of this summary is the calculation of the top sixth of the most searched items and sorted them in descending order. It is applied for non-numeric value.

```
       date     number_items_sold number_page_view sales_conversion  total_orders      sales_euro        returned        delivered
2017-01-01:  1   Min.   : 19.0    Min.   : 259.0   Min.   : 3.664   Min.   :  3.00   Min.   : 144.8   Min.   : 20.00   Min.   : 31.00
2017-01-02:  1   1st Qu.: 64.0    1st Qu.: 862.8   1st Qu.: 6.911   1st Qu.: 26.00   1st Qu.: 845.3   1st Qu.: 61.75   1st Qu.: 86.75
2017-01-03:  1   Median :108.0    Median :1417.0   Median : 7.944   Median : 44.50   Median :1518.7   Median :111.00   Median :156.50
2017-01-04:  1   Mean   :145.4    Mean   :1718.2   Mean   : 8.137   Mean   : 58.91   Mean   :1879.0   Mean   :136.18   Mean   :195.08
2017-01-05:  1   3rd Qu.:217.0    3rd Qu.:2325.5   3rd Qu.: 9.431   3rd Qu.: 84.25   3rd Qu.:2614.8   3rd Qu.:208.25   3rd Qu.:286.50
2017-01-06:  1   Max.   :423.0    Max.   :5252.0   Max.   :13.076   Max.   :215.00   Max.   :5490.9   Max.   :447.00   Max.   :625.00
(Other)   :206
 percreturned       most_searched_term total_num_term_searches  femalesales       malesales      femalesales_1945 femalesales_1965
Min.   :37.88   bikini dames  :99       Min.   : 10.00          Min.   :  3.00   Min.   : 0.000   Min.   :0        Min.   :0
1st Qu.:66.05   bikini        :80       1st Qu.: 21.00          1st Qu.: 23.00   1st Qu.: 2.000   1st Qu.:0        1st Qu.:0
Median :69.81   badpak        :16       Median : 29.50          Median : 41.00   Median : 5.000   Median :0        Median :0
Mean   :69.38   tankini       :11       Mean   : 37.66          Mean   : 52.06   Mean   : 6.844   Mean   :0        Mean   :0
3rd Qu.:73.08   bikini meisjes: 2       3rd Qu.: 50.00          3rd Qu.: 76.00   3rd Qu.:10.000   3rd Qu.:0        3rd Qu.:0
Max.   :93.88   badmode dames : 1       Max.   :118.00          Max.   :194.00   Max.   :29.000   Max.   :0        Max.   :0
                (Other)       : 3
femalesales_1985 femalesales_2005 femalesales_2017 malesales_1945 malesales_1965 malesales_1985   malesales_2005  malesales_2017       TG
Min.   : 0.000   Min.   : 0.00    Min.   :  1.00   Min.   :0      Min.   :0      Min.   :0.0000   Min.   :0.000   Min.   : 0.00   Min.   :-3.80
1st Qu.: 0.000   1st Qu.: 5.00    1st Qu.: 16.75   1st Qu.:0      1st Qu.:0      1st Qu.:0.0000   1st Qu.:0.000   1st Qu.: 2.00   1st Qu.: 6.10
Median : 1.000   Median :11.00    Median : 29.00   Median :0      Median :0      Median :0.0000   Median :0.000   Median : 4.00   Median : 9.65
Mean   : 1.307   Mean   :15.87    Mean   : 34.89   Mean   :0      Mean   :0      Mean   :0.2453   Mean   :1.269   Mean   : 5.33   Mean   :10.73
3rd Qu.: 2.000   3rd Qu.:23.00    3rd Qu.: 51.25   3rd Qu.:0      3rd Qu.:0      3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.: 7.00   3rd Qu.:16.73
Max.   :11.000   Max.   :54.00    Max.   :144.00   Max.   :0      Max.   :0      Max.   :5.0000   Max.   :9.000   Max.   :27.00   Max.   :23.60
```

Figure 1. Statistic Summary

Histograms

A histogram is an accurate graphical representation of the distribution of numerical data. The statistical information can be displayed using rectangles to show the frequency of data items in successive numerical intervals of equal size. Below histograms describe the distribution of data from each variable from January to July 2017.

- **Page Views**

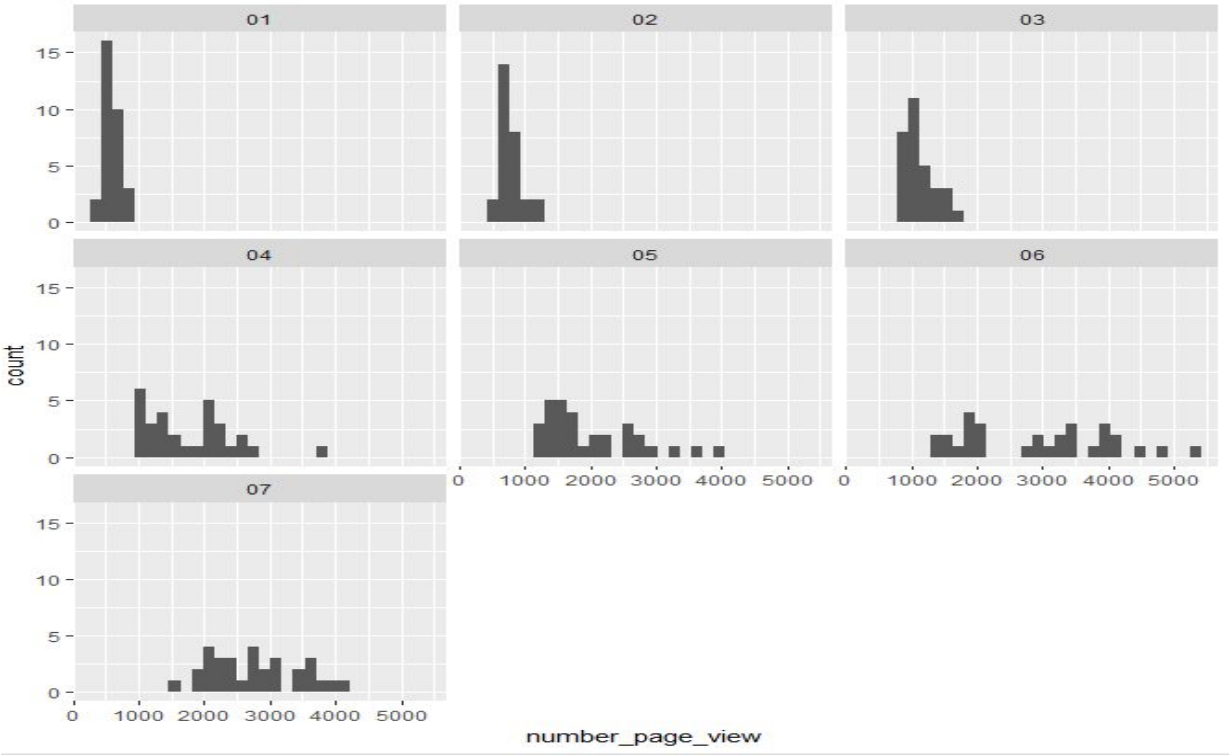

Figure 2. Histogram of number_page_view per month

The January and February data formed normal distribution with the highest peak of the total count is over 15. However, the value of the page view in these months are below 1500. In March, the data formed right skewed distribution with the highest peak is around 11 and the value moved up to almost 2000. In the following months, the value increases significantly in terms of page view but not the total count but they still formed a normal distribution.

- **Conversion Rate**

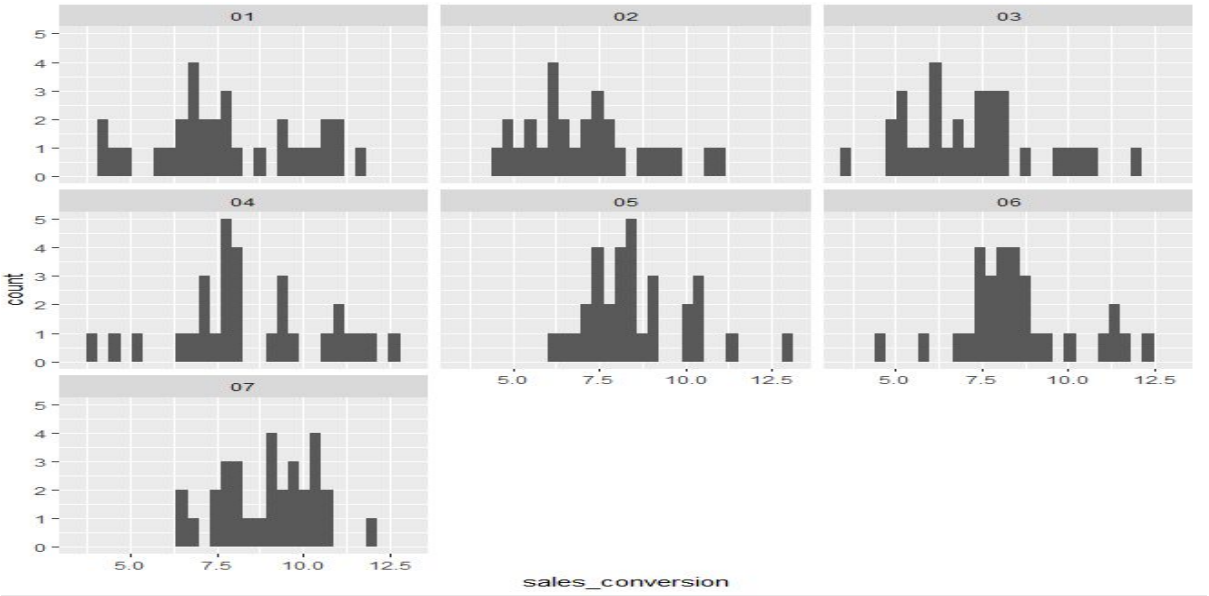

Figure 3. Histogram of conversion rate per month

From Figure 3, we can see that data of conversion rate is normally distributed. The mean values of conversion rate is 8 with the standard deviation is nearly 2.

- **Sales**

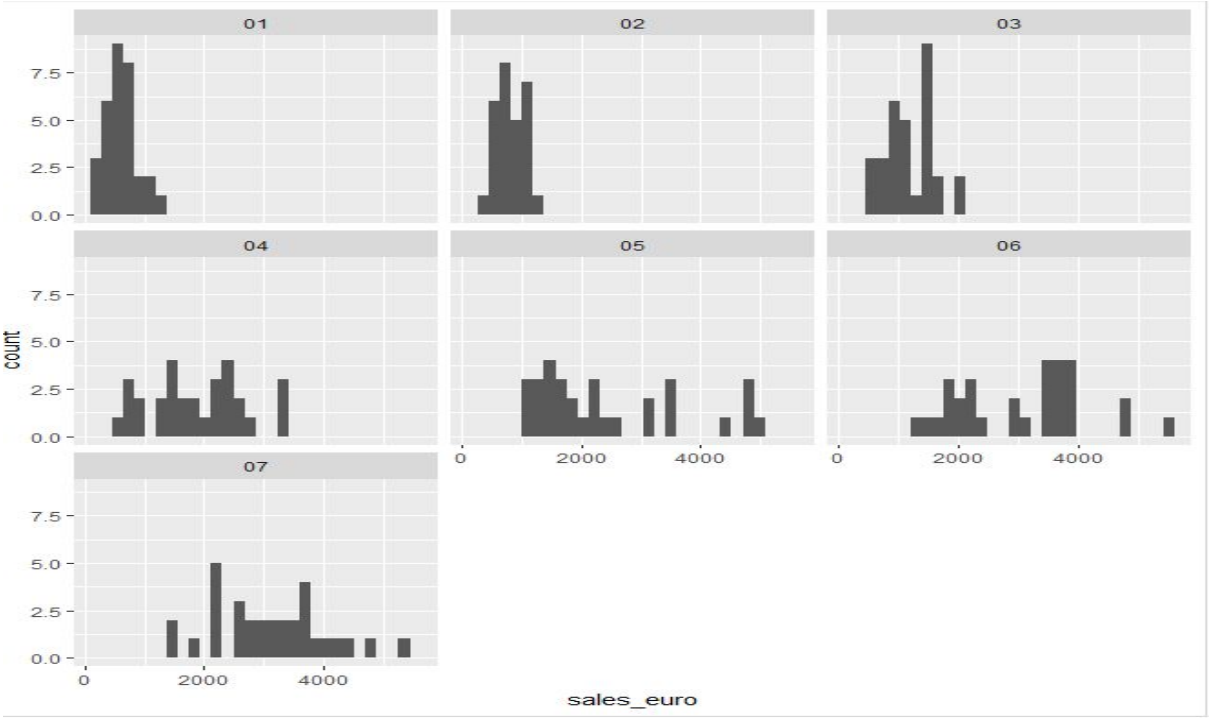

Figure 4. Histogram of sales per month

In January and February 2017, wekhamp does not have significant sales. However, this trend gradually changes in the following months. It reaches its sales up to 2000 starting from April 2017.

- **Delivered Item**

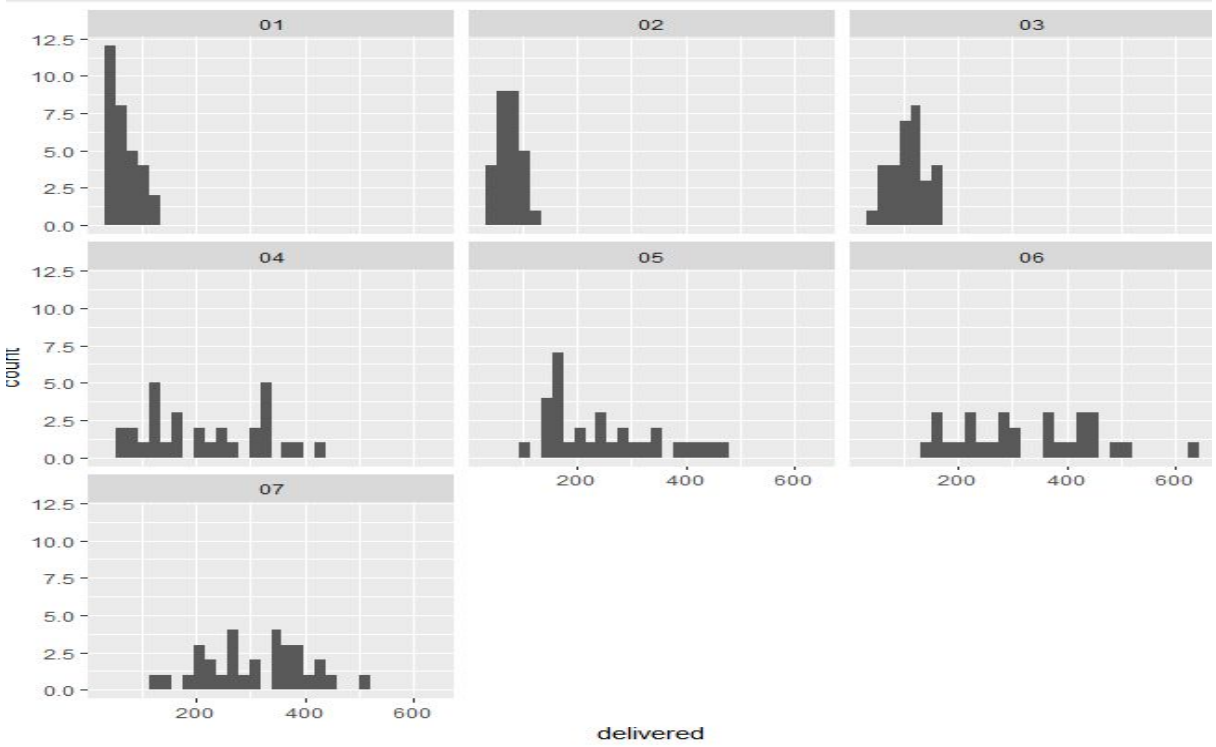

Figure 5. Histogram of delivered item per month

From Figure 5, it can be seen that the number of delivered items from January until March are not really significant. The average delivered items in these months is 85 and this number rises significantly in the following months by reaching the average of 195 delivered items.

- **Returned Item**

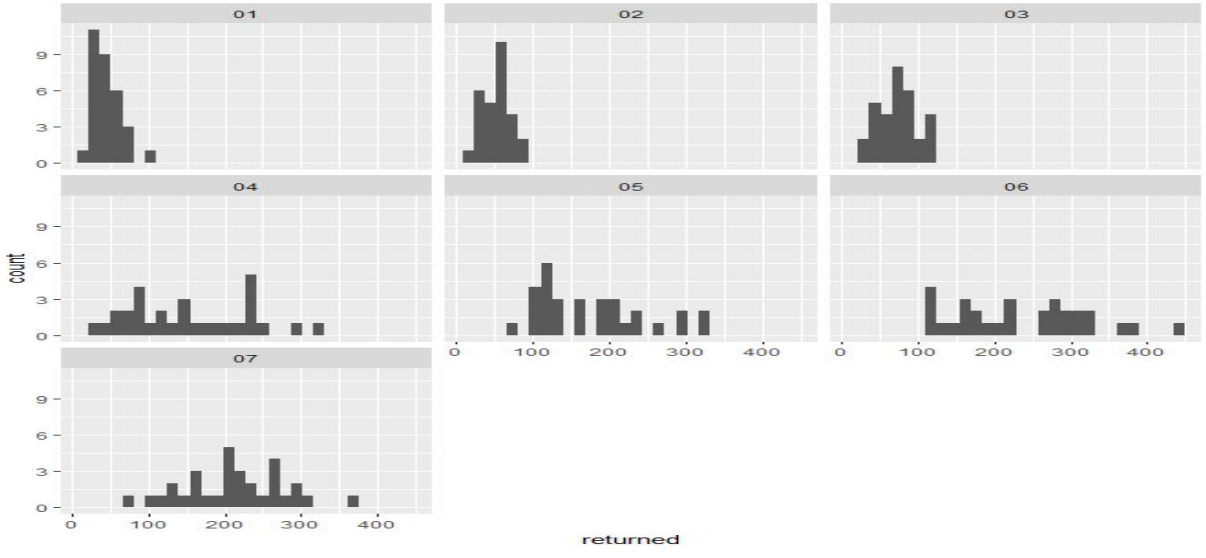

Figure 6. Histogram of returned item per month

The trend of returned item is almost similar with delivered item. We are not saying that they have similar number in terms of quantity, but the curve of the histograms are alike as depicted in Figure 6. The average of returned items in January,February, and March is 58. This number significantly rises by 136 in the following months.

- **Male Sales**

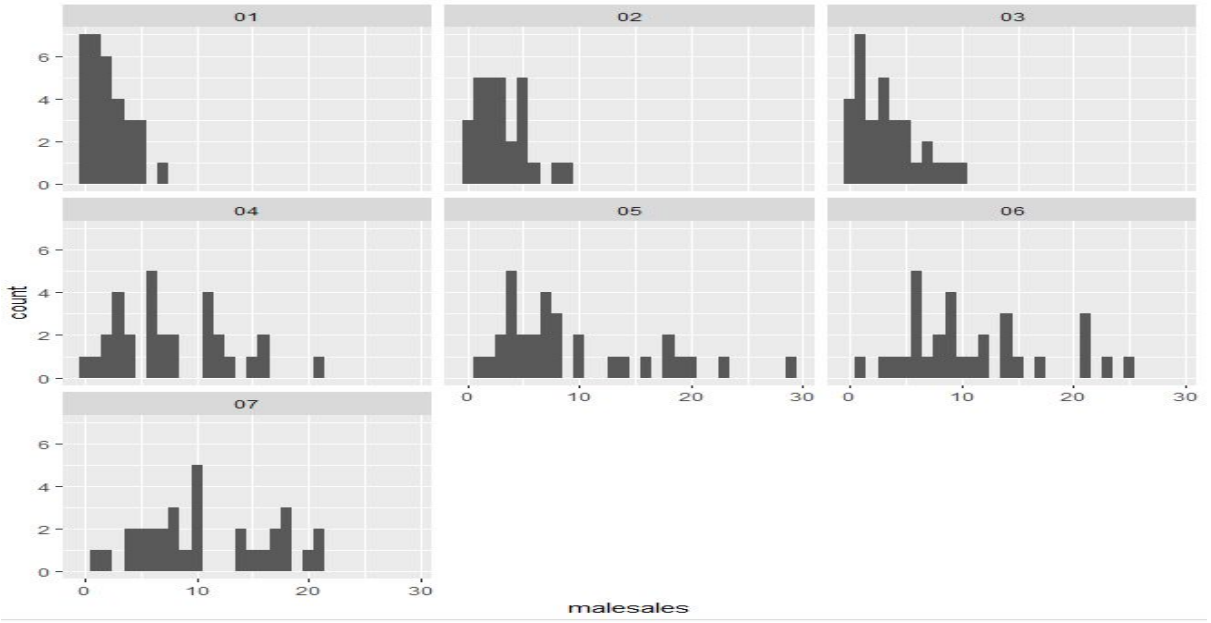

Figure 7. Histogram of overall sold orders made by male customers per month

Figure 7 depicts the histogram of overall sales made by male customers. In the first three month, the average sold order is 2 and this number gradually increases to 7 in the last four months. Data from April to July forms normal distribution with standard deviation is 6.

The details of the orders made by males customer are described in the histograms in figure 8, 9, 10, 11 and 12 respectively. From these histograms we can draw a conclusion that the male customers above 31 never did shopping on wekhamp.

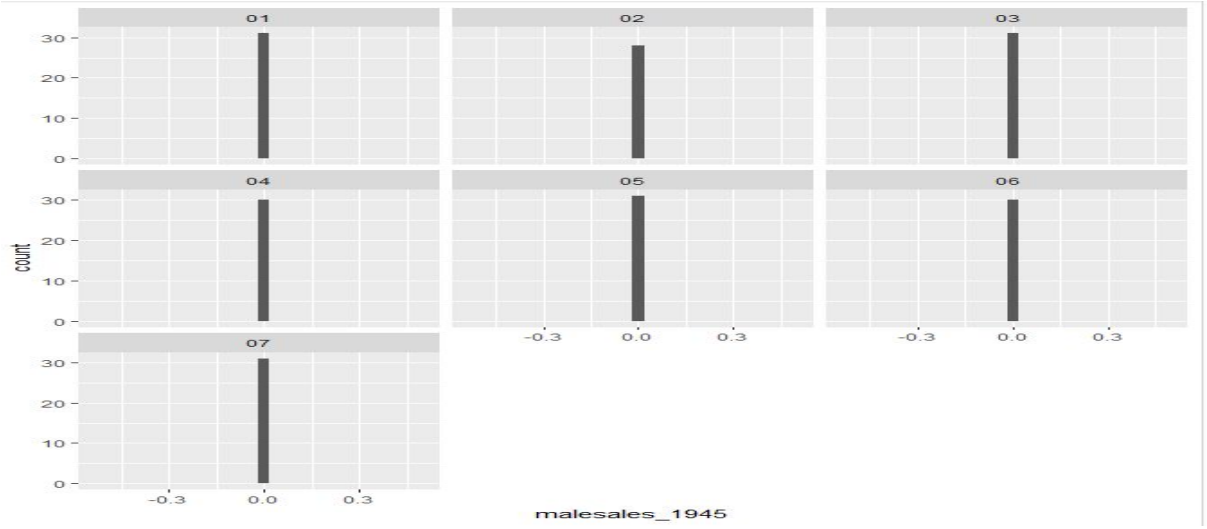

Figure 8. Histogram of male customers sales born between 1900 and 1945 per month

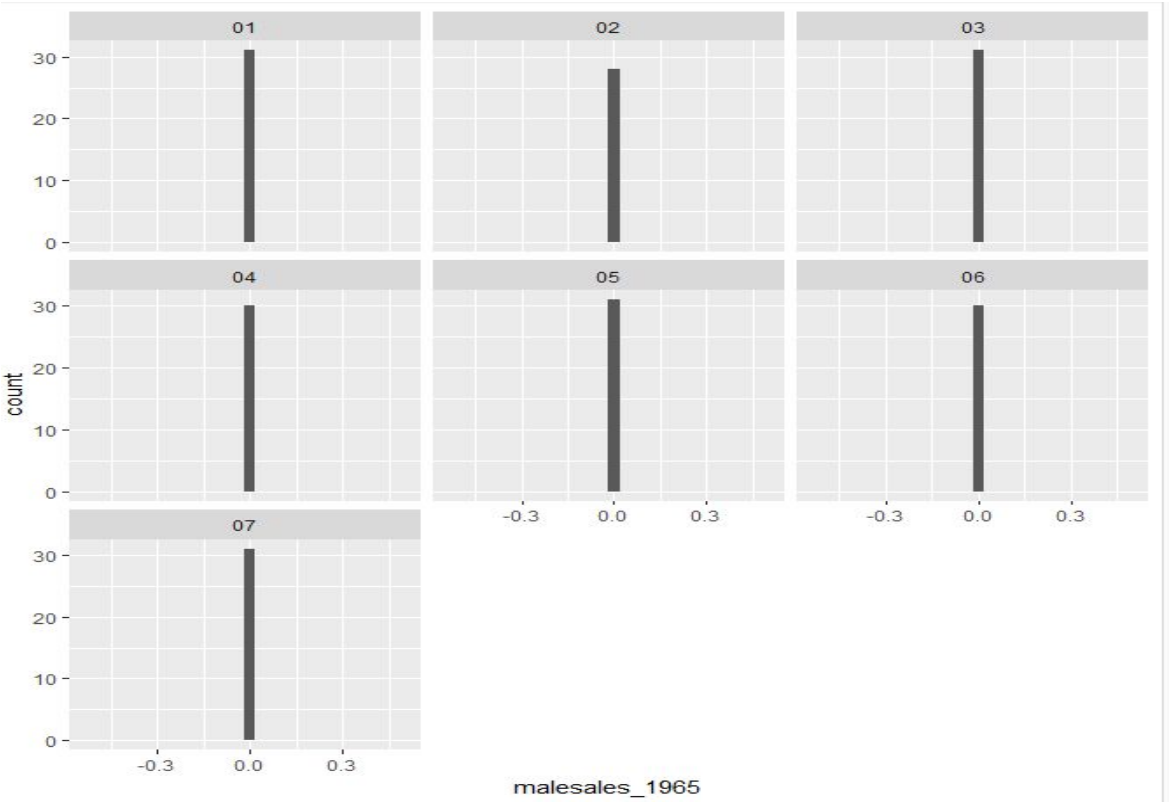

Figure 9. Histogram of male customers sales born between 1946 and 1965 per month

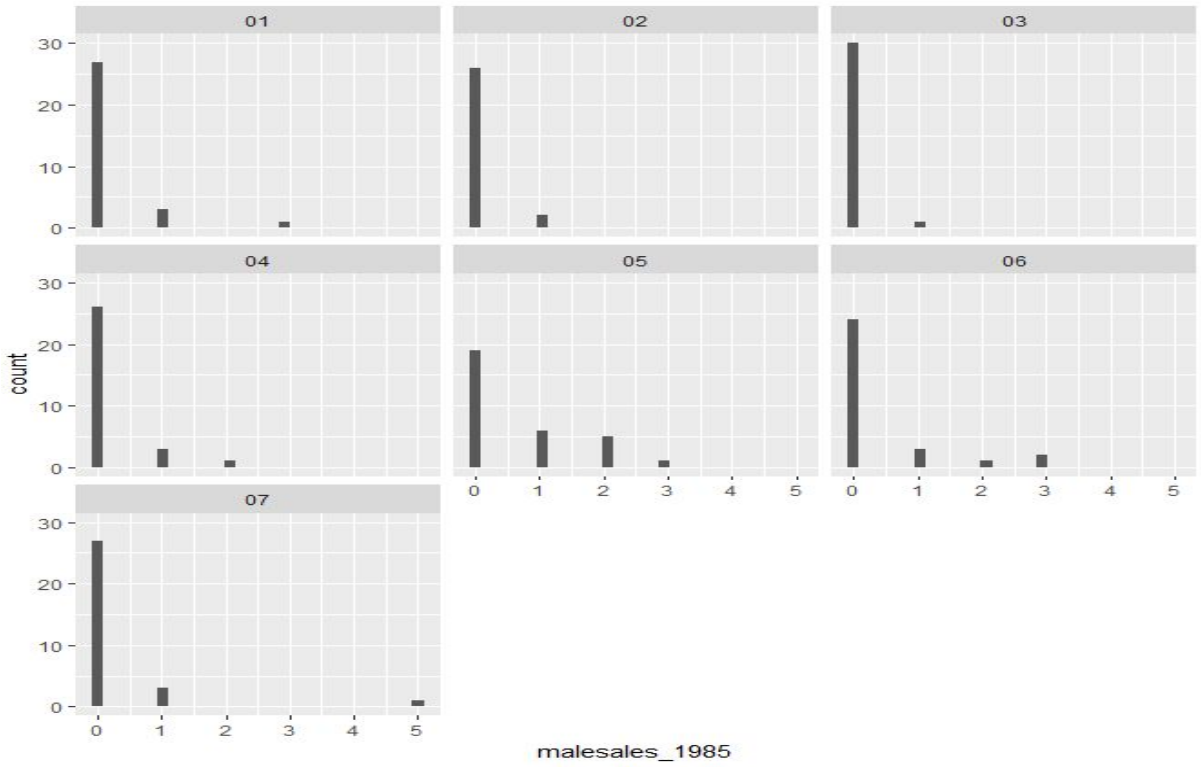

Figure 10. Histogram of male customers sales born between 1966 and 1985 per month

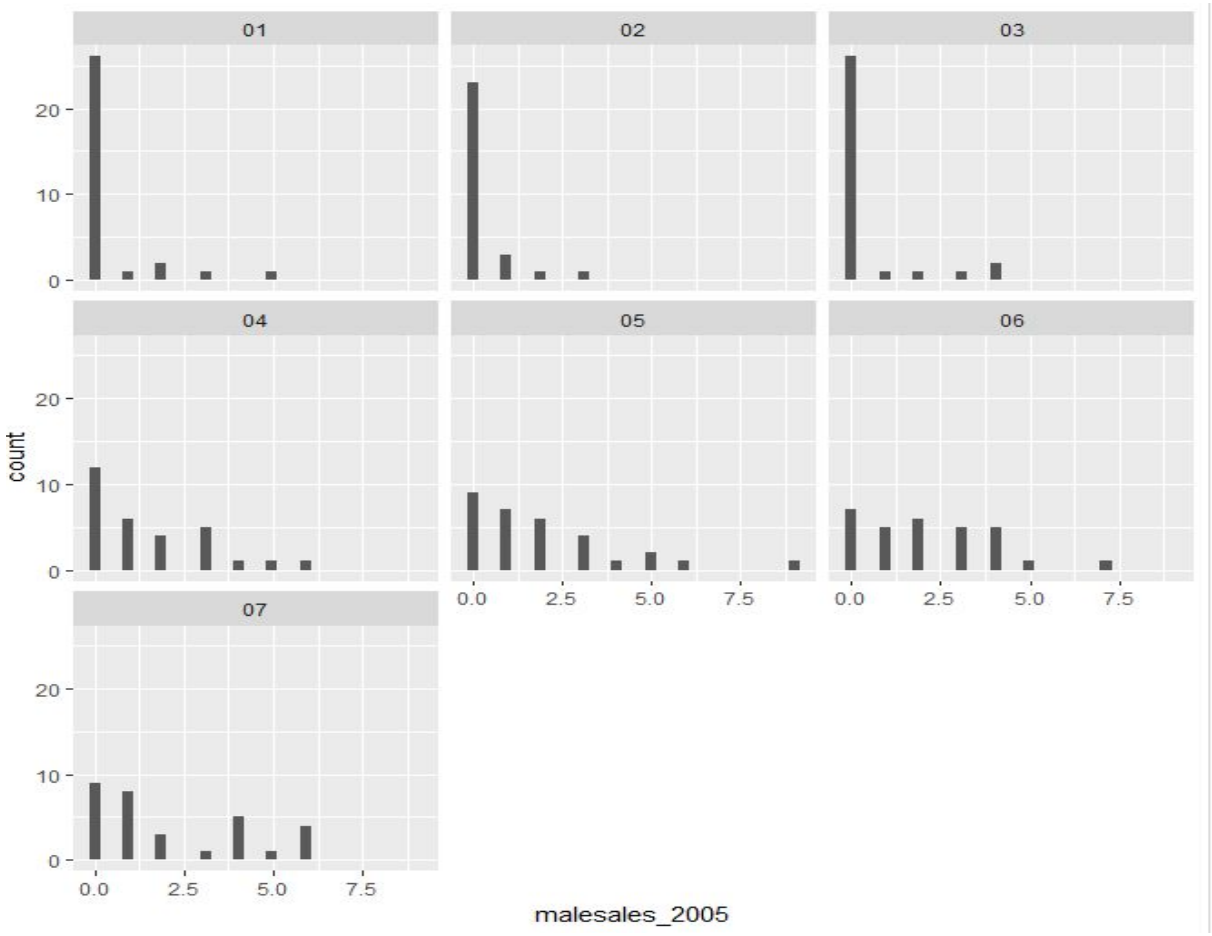

Figure 11. Histogram of male customers sales born between 1986 and 2005 per month

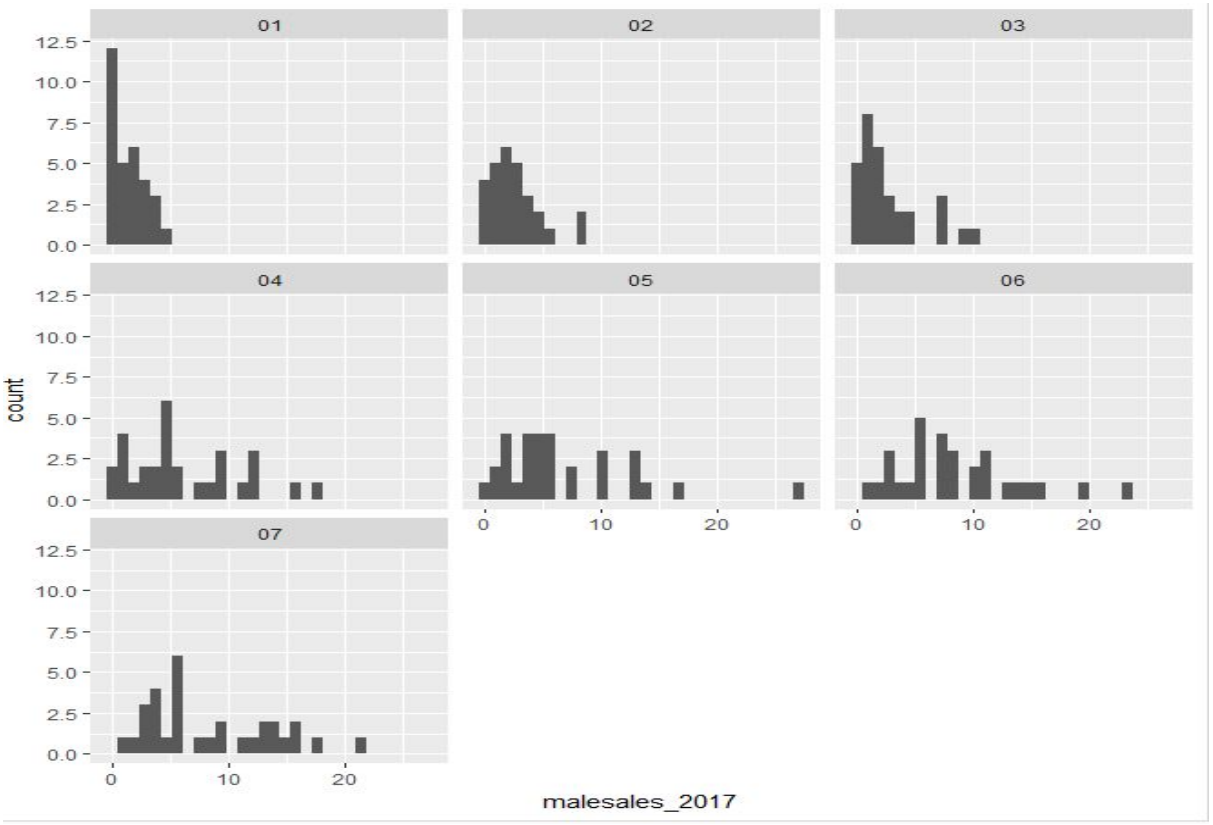

Figure 12. Histogram of male customers sales born between 2006 and 2017 per month

- **Female Sales**

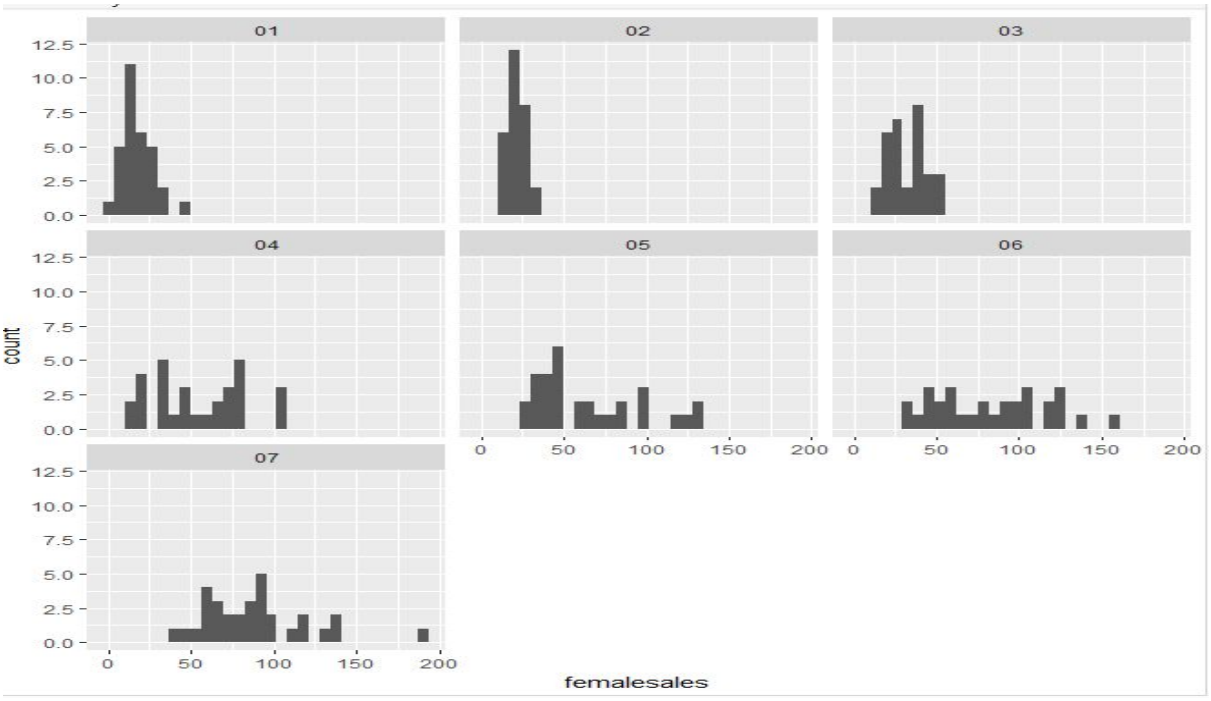

Figure 13. Histogram of overall orders made by female customers per month

Figure 13 depicts the histogram of overall sales made by female customers. In the first three month, the average sold order is 23 and this number gradually increases

to 52 in the last four months. This number is ten times higher than the orders made by male customers.

The details of the orders made by females customer are described in the histograms in figure 14, 15, 16, 17 and 18 respectively. From these histograms, we can notice that the female customers who are above 51 never bought products from wehkamp. Conversely, the female customers below 31 tend to by products more often on the website.

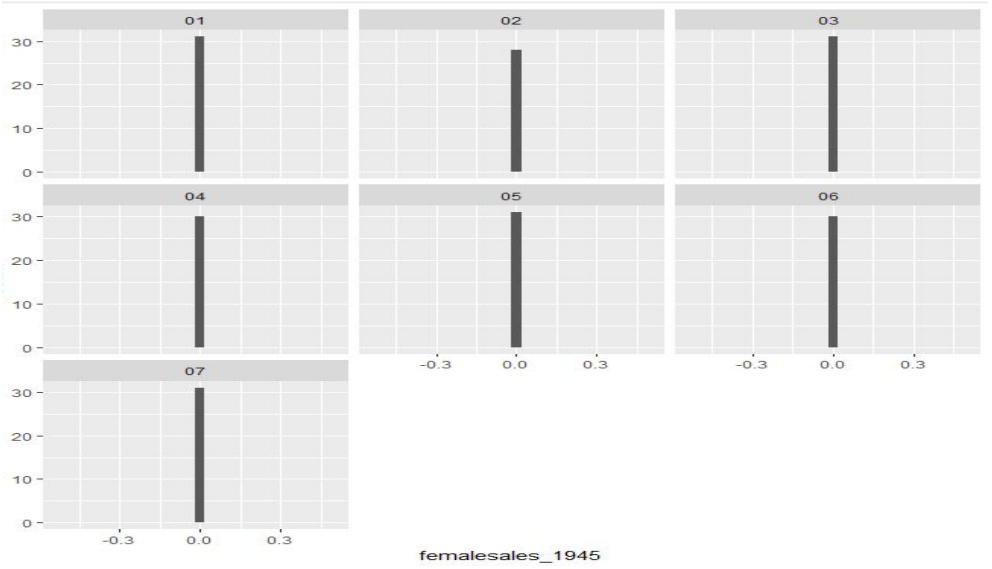

Figure 14. Histogram of female customers sales born between 1900 and 1945 per month

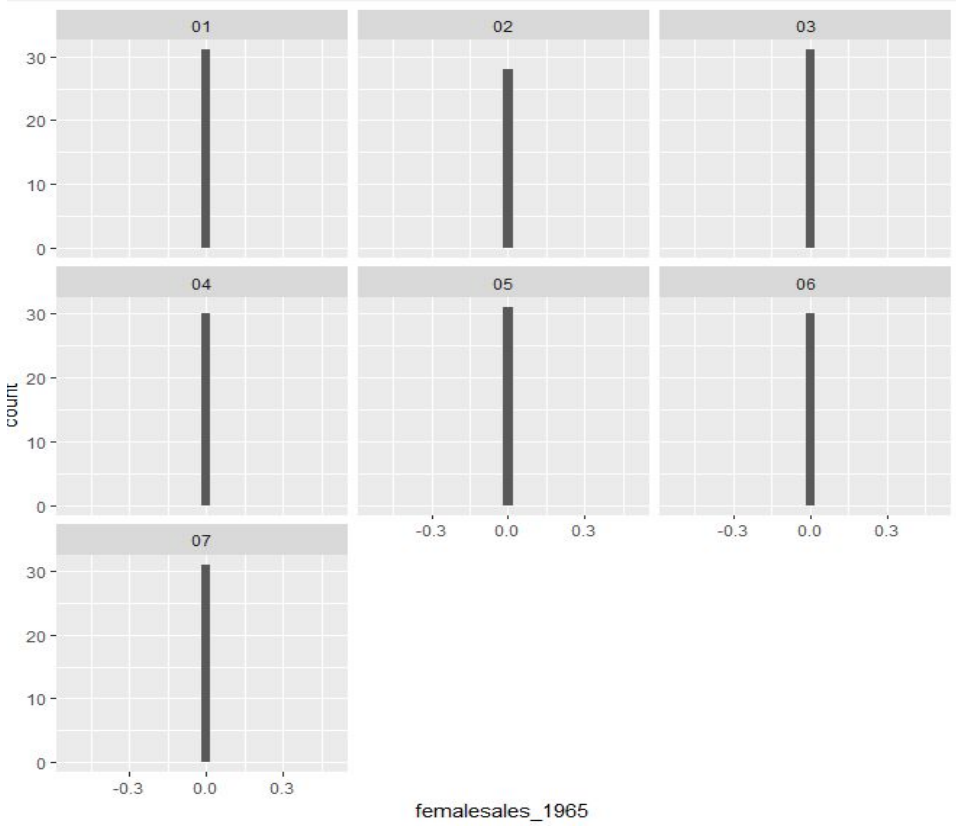

Figure 15. Histogram of female customers sales born between 1946 and 1965 per month

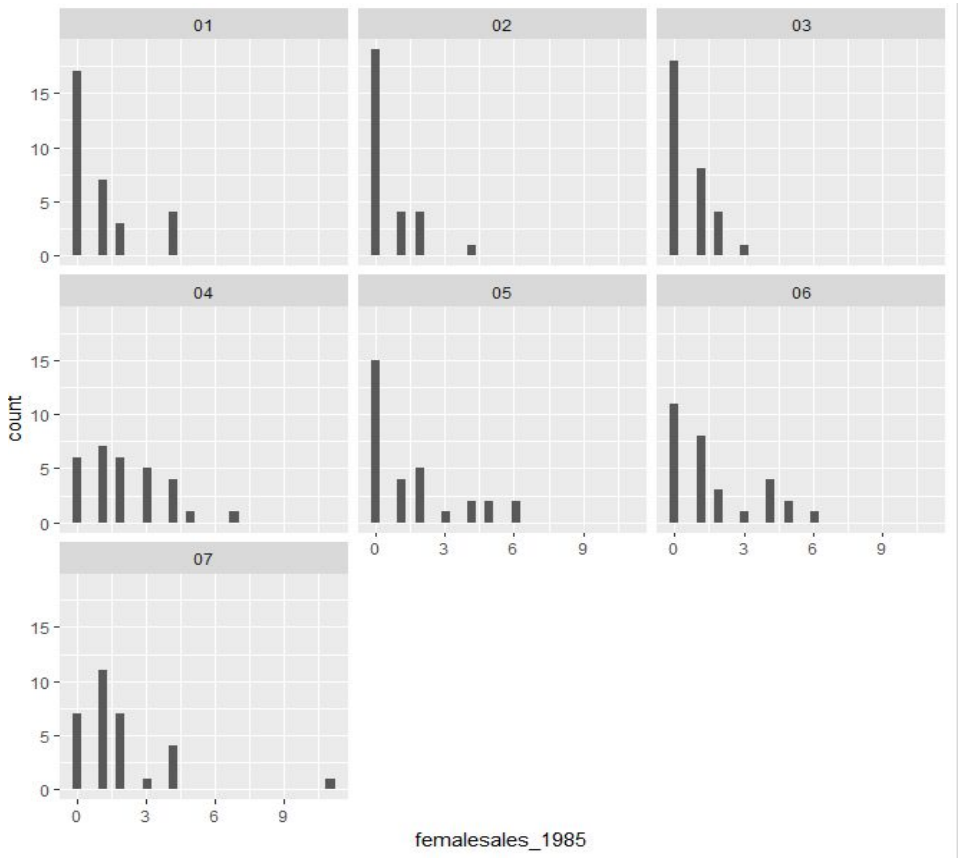

Figure 16. Histogram of female customers sales born between 1966 and 1985 per month

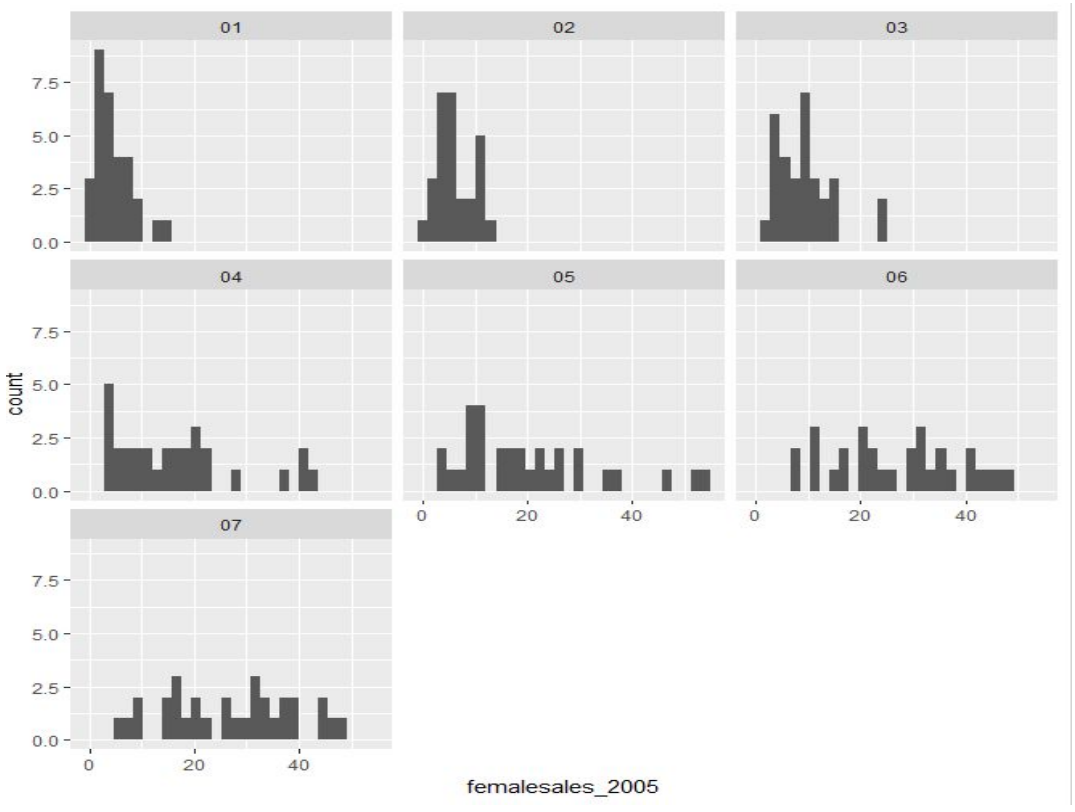

Figure 17. Histogram of female customers sales born between 1986 and 2005 per month

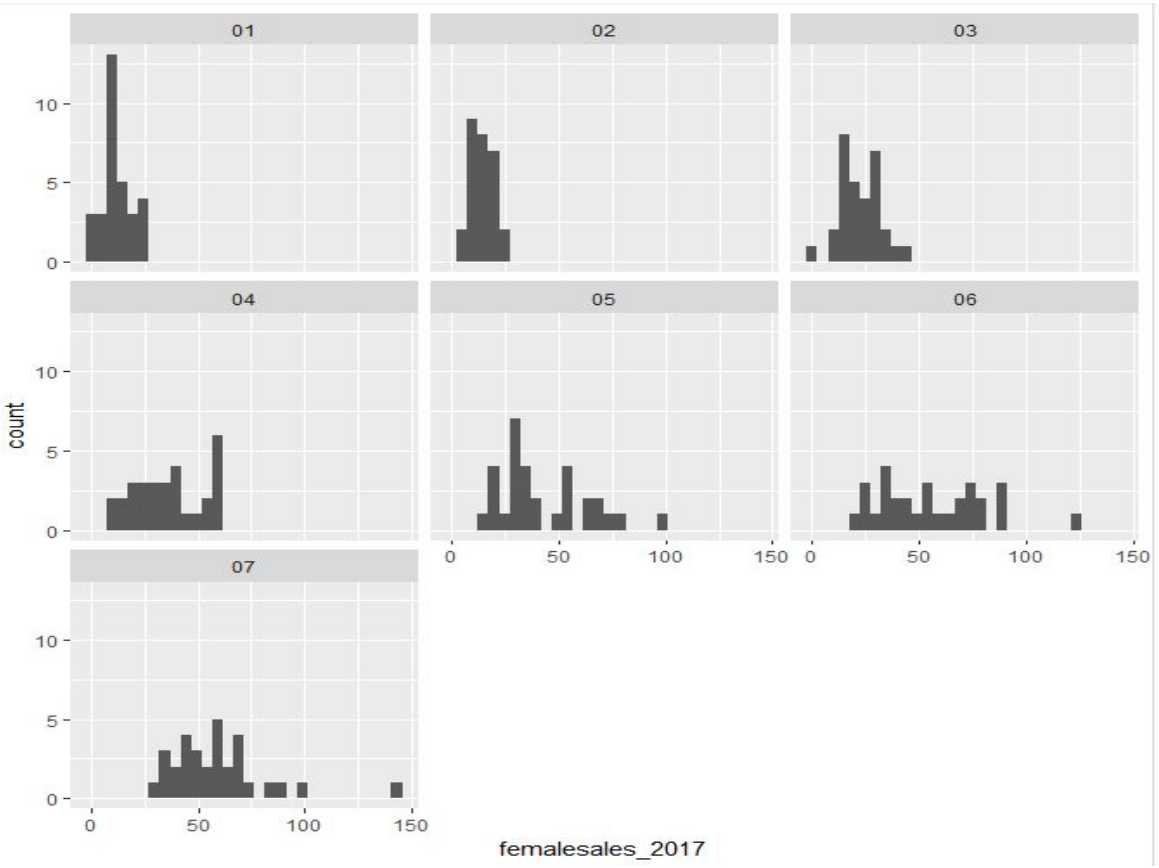

Figure 18. Histogram of female customers sales born between 2006 and 2017 per month

- **Temperature**

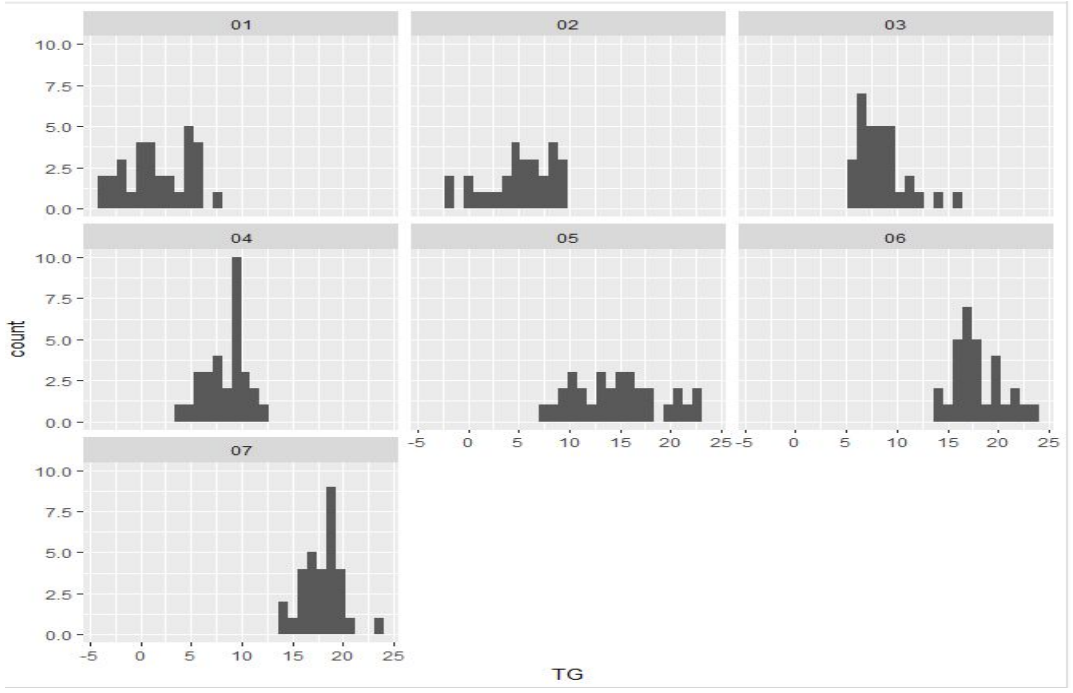

Figure 19. Histogram of temperature per month

Figure 19 depicts the monthly average temperature in Netherlands. Almost all temperature data per month formed normal distributions. From the above

histogram we can see that the average temperature in January was 1.5 degrees. The temperature increased to 5 degrees in February. In the following month, it rose to 8.5 and was stable until April. Then, it rose again up to 17 by July.

- **Most Searched Items**

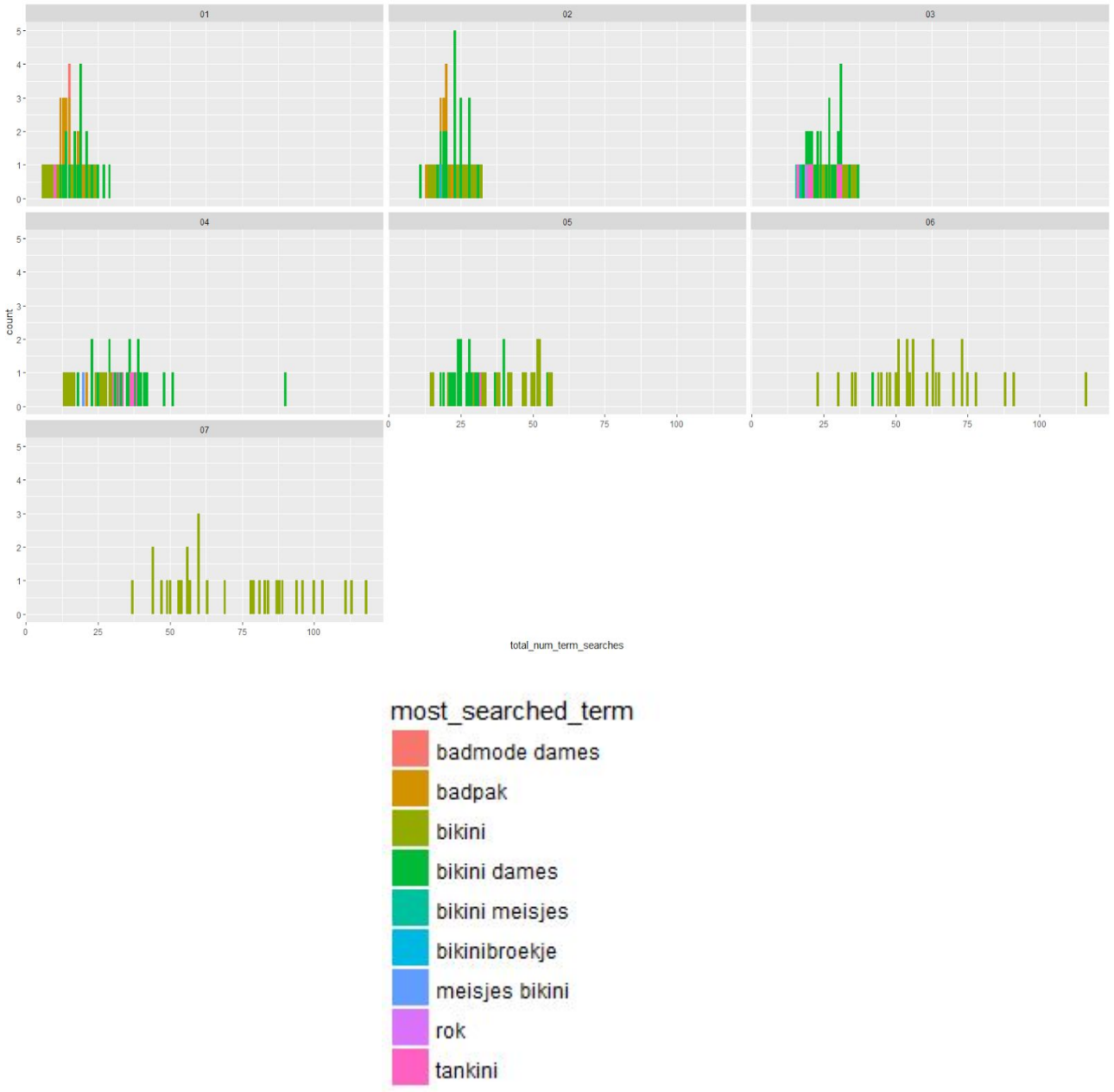

Figure 20. Histogram of the most searched item per-month

From the above figure, it can be seen that the number of frequencies of people searching for particular items on whekhamp website using specific keywords is not more than 50 time per day starting from January until May. This trend gradually increases in the following months, reaching up to 180 times per day by July 2017.

Being the most searched item, the average number of bikini for the last seven months is 58 followed by bikini dames at 28, tankini 27, meijes bikini 20, bikini meisjes 19, rok 19, bikinibroekje 18, badpak 16 , and badmode dames 15.

Missing Value in Dataset

In order to validate the completion of dataset, we performed a test using mice library. However, the result shows nothing are missing from our dataset. The source code for validating the missing value can be found in appendix.

APPENDIX (SQL QUERY AND R CODE)

Impression / Pageview SQL query :

```
SELECT
article_event_date,
COUNT (CASE WHEN article_event_type='10' THEN 1 ELSE NULL END) AS number_views
FROM public.articleevents
GROUP BY article_event_date
ORDER BY article_event_date
```

Conversion SQL query :

```
SELECT *,
ROUND(number_sales*100.0/number_views,3) AS sales_conversion
FROM
(
SELECT
article_event_date,
COUNT(CASE WHEN article_event_type='40' THEN 1 ELSE NULL END) AS number_sales,
COUNT (CASE WHEN article_event_type='10' THEN 1 ELSE NULL END) AS number_views
FROM public.articleevents
GROUP BY article_event_date
ORDER BY article_event_date
) AS conversion
```

Sales SQL Query :

```
select
b.order_date,
sum(b.items) as total_items,
sum(b.sales_amount) as total_sales
from public.article a
inner join public.order b on a.article_id = b.article_id
where a.category = 'Beachwear' and b.delivered = 1 and b.returned = 0
group by b.order_date
order by b.order_date asc
```

Most Searched Item SQL query :

```
WITH cte as (
SELECT *,
ROW_NUMBER() OVER (PARTITION BY onsite_search_date) AS rn
FROM
(
SELECT
search_term,
max(most_search),
onsite_search_date
FROM
(
SELECT
search_term,
COUNT(search_term) AS most_search,
onsite_search_date
FROM public.onsitesearches
GROUP BY search_term, onsite_search_date
ORDER BY onsite_search_date, most_search DESC
, search_term
) AS terms
GROUP BY onsite_search_date, search_term
ORDER BY onsite_search_date asc, max desc
) as maxcount
)
SELECT *
FROM cte
WHERE rn=1
```

Percentage of Returned Items SQL query:

```
SELECT order_date,
COUNT(CASE WHEN returned = 1 THEN 1 ELSE NULL END) AS Returned,
COUNT(CASE WHEN delivered = 1 THEN 1 ELSE NULL END) AS Delivered,
CAST(100.0 * COUNT(CASE WHEN returned = 1 THEN 1 ELSE NULL END) / COUNT(CASE WHEN
delivered = 1 THEN 1 ELSE NULL END) AS DECIMAL(4,2)) AS PercReturned
FROM public.order
GROUP BY order_date
ORDER BY order_date
)AS returns
ON returns.order_date=conversion.article_event_date
```

Final Dataset SQL query :

```
SELECT article_event_date AS date, number_sales AS number_items_sold, number_views AS
number_page_view, sales_conversion, total_items AS total_orders, total_sales AS
sales_euro, Returned, Delivered, PercReturned, search_term AS most_searched_term, max AS
total_num_term_searches, FemaleSales, MaleSales,
FemaleSales_1945, FemaleSales_1965, FemaleSales_1985, FemaleSales_2005,FemaleSales_2017,
```

```
MaleSales_1945, MaleSales_1965, MaleSales_1985, MaleSales_2005,MaleSales_2017
FROM
(
SELECT *, ROUND(number_sales*100.0/number_views,3) AS sales_conversion
FROM
(SELECT
article_event_date,
COUNT(CASE WHEN article_event_type='40' THEN 1 ELSE NULL END) AS number_sales,
COUNT (CASE WHEN article_event_type='10' THEN 1 ELSE NULL END) AS number_views
FROM public.articleevents
GROUP BY article_event_date
ORDER BY article_event_date) AS conversion
)as conversion
LEFT JOIN
(
select b.order_date, sum(b.items) as total_items, sum(b.sales_amount) as total_sales
from public.article a inner join public.order b on a.article_id = b.article_id
where a.category = 'Beachwear' and b.delivered = 1 and b.returned = 0 group by
b.order_date order by b.order_date asc
)as sales
ON conversion.article_event_date=sales.order_date
LEFT JOIN
(
SELECT order_date,
COUNT(CASE WHEN returned = 1 THEN 1 ELSE NULL END) AS Returned,
COUNT(CASE WHEN delivered = 1 THEN 1 ELSE NULL END) AS Delivered,
CAST(100.0 * COUNT(CASE WHEN returned = 1 THEN 1 ELSE NULL END) / COUNT(CASE WHEN
delivered = 1 THEN 1 ELSE NULL END) AS DECIMAL(4,2)) AS PercReturned
FROM public.order
GROUP BY order_date
ORDER BY order_date
)AS returns
ON returns.order_date=conversion.article_event_date

LEFT JOIN

(
WITH cte as (
SELECT *,
ROW_NUMBER() OVER (PARTITION BY onsite_search_date) AS rn
FROM
(
SELECT search_term, max(most_search), onsite_search_date
FROM
(
SELECT search_term, COUNT(search_term) AS most_search, onsite_search_date
FROM public.onsitesearches
GROUP BY search_term, onsite_search_date
ORDER BY onsite_search_date, most_search DESC
, search_term ) AS terms
GROUP BY onsite_search_date, search_term
ORDER BY onsite_search_date asc, max desc
) as maxcount
```

```
)
SELECT *
FROM cte
WHERE rn=1
) AS mostsearched

ON mostsearched.onsite_search_date=conversion.article_event_date

LEFT JOIN
(
SELECT order_date,
COUNT(CASE
WHEN customer.sex='F' THEN 1 ELSE NULL END) AS FemaleSales,
COUNT(CASE
WHEN customer.sex='M' THEN 1 ELSE NULL END) AS MaleSales,
COUNT(*) AS TotalSales,
COUNT(CASE
WHEN customer.sex='F' AND customer.start_year BETWEEN '1900' AND '1945' THEN 1 ELSE NULL
END) AS FemaleSales_1945,
COUNT(CASE
WHEN customer.sex='F' AND customer.start_year BETWEEN '1946' AND '1965' THEN 1 ELSE NULL
END) AS FemaleSales_1965,
COUNT(CASE
WHEN customer.sex='F' AND customer.start_year BETWEEN '1966' AND '1985' THEN 1 ELSE NULL
END) AS FemaleSales_1985,
COUNT(CASE
WHEN customer.sex='F' AND customer.start_year BETWEEN '1986' AND '2005' THEN 1 ELSE NULL
END) AS FemaleSales_2005,
COUNT(CASE
WHEN customer.sex='F' AND customer.start_year BETWEEN '2006' AND '2017' THEN 1 ELSE NULL
END) AS FemaleSales_2017,
COUNT(CASE
WHEN customer.sex='M' AND customer.start_year BETWEEN '1900' AND '1945' THEN 1 ELSE NULL
END) AS MaleSales_1945,
COUNT(CASE
WHEN customer.sex='M' AND customer.start_year BETWEEN '1946' AND '1965' THEN 1 ELSE NULL
END) AS MaleSales_1965,
COUNT(CASE
WHEN customer.sex='M' AND customer.start_year BETWEEN '1966' AND '1985' THEN 1 ELSE NULL
END) AS MaleSales_1985,
COUNT(CASE
WHEN customer.sex='M' AND customer.start_year BETWEEN '1986' AND '2005' THEN 1 ELSE NULL
END) AS MaleSales_2005,
COUNT(CASE
WHEN customer.sex='M' AND customer.start_year BETWEEN '2006' AND '2017' THEN 1 ELSE NULL
END) AS MaleSales_2017
FROM public.order, public.customer
WHERE customer.customer_id=public.order.customer_id AND public.order.delivered = 1 and
public.order.returned = 0
GROUP BY order_date
ORDER BY order_date
) AS gender
ON conversion.article_event_date=gender.order_date
```

R code for merging temperature and sql queries together:

```
rm(list = ls()) #clear workspace

setwd("x://My Downloads/")  #set working directory

wehkamp_data <- read.csv("sql_kpi.csv",header=TRUE) #read csv file

wehkamp_data$date <- as.Date(wehkamp_data$date) #format date field

summary (wehkamp_data)

temperature_data <- read.csv("Temperature.csv",header=TRUE) #read csv file

temperature_data$date <- as.Date(as.factor(temperature_data$YYYYMMDD),"%Y%m%d") #give a
few more options

required_fields <- as.vector(c("date","TG")) #obtain only date and temperature fields

temperature_subdata <- temperature_data[,required_fields]

# temperature_subdata$date <- as.Date(temperature_subdata$date)

temperature_data <- with(temperature_subdata, temperature_subdata[(date >= "2017-01-01"
& date <= "2017-07-31"), ]) #filter only dates of interest
temperature_data$TG <- temperature_data$TG/10; #convert temperature to a normal scale
summary(temperature_data)

final_dataset <- merge(wehkamp_data, temperature_data) #merge the two datasets

write.csv(final_dataset, file = "Assignment1_dataset.csv",row.names=TRUE) #Save the
final dataset as csv file
```

R code for description of variables

```
library(ggplot2)

#load final dataset
df <- read.csv(file="Assignment1_dataset.csv")
df <- subset(df, select = -c(X) )
df$date <- as.Date(as.factor(df$date),"%Y-%m-%d")
df$month <- format(df$date,"%m")

#get summary of dataframe
df_summary <- summary(df)

#histogram

#number_items_sold number_page_view sales_conversion  total_orders
sales_euro        returned        delivered
```

```
#percreturned        most_searched_term total_num_term_searches  femalesales
malesales      femalesales_1945 femalesales_1965
#femalesales_1985 femalesales_2005 femalesales_2017 malesales_1945
malesales_1965 malesales_1985   malesales_2005  malesales_2017        TG

#number_page_view
ggplot(df, aes(x=number_page_view)) + geom_histogram() + facet_wrap(~month)

#sales_conversion
ggplot(df, aes(x=sales_conversion)) + geom_histogram() + facet_wrap(~month)

#sales_euro
ggplot(df, aes(x=sales_euro)) + geom_histogram() + facet_wrap(~month)

#delivered
ggplot(df, aes(x=delivered)) + geom_histogram() + facet_wrap(~month)

#returned
ggplot(df, aes(x=returned)) + geom_histogram() + facet_wrap(~month)

#femalesales
ggplot(df, aes(x=femalesales)) + geom_histogram() + facet_wrap(~month)

#malesales
ggplot(df, aes(x=malesales)) + geom_histogram() + facet_wrap(~month)

#femalesales_1945
ggplot(df, aes(x=femalesales_1945)) + geom_histogram() + facet_wrap(~month)

#femalesales_1965
ggplot(df, aes(x=femalesales_1965)) + geom_histogram() + facet_wrap(~month)

#femalesales_1985
ggplot(df, aes(x=femalesales_1985)) + geom_histogram() + facet_wrap(~month)

#femalesales_2005
ggplot(df, aes(x=femalesales_2005)) + geom_histogram() + facet_wrap(~month)

#femalesales_2017
ggplot(df, aes(x=femalesales_2017)) + geom_histogram() + facet_wrap(~month)

#malesales_1945
ggplot(df, aes(x=malesales_1945)) + geom_histogram() + facet_wrap(~month)

#malesales_1965
ggplot(df, aes(x=malesales_1965)) + geom_histogram() + facet_wrap(~month)

#malesales_1985
ggplot(df, aes(x=malesales_1985)) + geom_histogram() + facet_wrap(~month)

#malesales_2005
ggplot(df, aes(x=malesales_2005)) + geom_histogram() + facet_wrap(~month)

#malesales_2017
ggplot(df, aes(x=malesales_2017)) + geom_histogram() + facet_wrap(~month)

#temperature
```

```
ggplot(df, aes(x=TG)) + geom_histogram() + facet_wrap(~month)

#most_searched_term
ggplot(df, aes(x=total_num_term_searches, fill=most_searched_term)) + geom_bar()
+ facet_wrap(~month)
```

R code for validating the completion of the dataset:

```
#use MICE library
library(mice)

#inspect pattern of missings
md.pattern(final_dataset)
```

References

Bhapkar, Neil, (2013) "8 KPIs Your Content Marketing Measurement Should Include", Content Marketing Institute. Retrieved from: http://contentmarketinginstitute.com/2013/02/kpis-for-content-marketing-measurement/.

DeMers, Jayson, (2014) "10 Online Marketing Metrics You Need To Be Measuring", Forbes. Retrieved from: https://www.forbes.com/sites/jaysondemers/2014/08/15/10-online-marketing-metrics-you-need-to-be-measuring/#371a6e1376c1.

Gagandeep Nagra, R. Gopal, (2013) "A study of Factors Affecting on Online Shopping Behavior of Consumers", International Journal of Scientific and Research Publications, Vol. 3, Issue: 6.

Markert, J. (2004). Demographics of Age: Generational and Cohort Confusion. Journal of Current Issues & Research in Advertising, 26(2), 11–25.

Mummalaneni, Venkatapparao, (2005) "An empirical investigation of Web site characteristics, consumer emotional states and on-line shopping behaviors", Journal of Business Research, Vol. 58, pp. 526 – 532.

Lazazzera, Richard, (2014), "The Beginner's Guide To Keyword Research For Ecommerce", Shopify. Retrieved from: https://www.shopify.com/blog/14207073-the-beginners-guide-to-keyword-research-for-ecommerce.

Parson, (2001) "The Association Between Daily Weather and Daily Shopping Patterns", Australasian Marketing Journal Vol. 9 Issue: 2, pp. 78-84.

Youngjin Bahng, Doris H. Kincade, (2012) "The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear", International Journal of Retail & Distribution Management, Vol. 40 Issue: 6, pp.410-426.