ACADGILD

SESSION 13: Decision Tree Based Models

1. Problem Statement

1. Use the given link below:
https://archive.ics.uci.edu/ml/machine-learning-databases/00304/

Problem- prediction of the number of comments in the upcoming 24 hours on those blogs, the train data was generated from different base times that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, the you should use the provided, temporally disjoint train and test splits to ensure that the evaluation is fair.

a) Read the dataset and identify the right features.
b) Clean dataset, impute missing values and perform exploratory data analysis.
c) Visualize the dataset and make inferences from that.
d) Perform any 3 hypothesis tests using columns of your choice, make conclusions.

2. Solution

a. Read the dataset and identify the right features.

The R-script for the given problem is as follows:

```
library(data.table)
library(foreach)
library(readr)
library(dplyr)

setwd("F:/ACADGILD - Online Course/1. DATA SETS/BlogFeedback")
getwd()

blogData_train <- read_csv("F:/ACADGILD - Online Course/1. DATA
SETS/blogData_train.csv")
View(blogData_train)

# retrieve filenames of test sets
test_filenames = list.files(pattern = "blogData_test")
```

```
# load and combine dataset
train = fread("blogData_train.csv")
fbtest = foreach(i = 1:length(test_filenames), .combine = rbind) %do%
  { temp = fread(test_filenames[i], header = F)
}

# Assign variable names to the train data set

 colnames(blogData_train) <-

c("plikes","checkin","talking","category","d5","d6","d7","d8","d9","d10","d11","d12",

"d13","d14","d15","d16","d17","d18","d19","d20","d21","d22","d23","d24","d25","d26",

"d27","d28","d29","cc1","cc2","cc3","cc4","cc5","basetime","postlength","postshre",

"postpromo","Hhrs","sun","mon","tue","wed","thu","fri","sat","basesun","basemon",
                  "basetue","basewed","basethu","basefri","basesat","target")

dim(blogData_train)
dim(fbtest)
View(blogData_train)
View(fbtest)
str(blogData_train)
str(fbtest)

train <- blogData_train; test <- fbtest
head(train); head(test)

# making the data tidy by constructing single collumn for post publish day
train$pubday<- ifelse(train$sun ==1, 1, ifelse(train$mon ==1, 2, ifelse(train$tue ==1,
                                                    3, ifelse(train$wed ==1, 4, ifelse(train$thu
==1, 5, ifelse(train$fri ==1, 6,

ifelse(train$sat ==1, 7, NA)))))))

# making the data tidy by constructing single collumn for base day
train$baseday<- ifelse(train$basesun ==1, 1, ifelse(train$basemon ==1,
2, ifelse(train$basetue ==1, 3,
                                                          ifelse(train$basewed ==1, 4,
ifelse(train$basethu ==1, 5,

ifelse(train$basefri ==1, 6, ifelse(train$basesat ==1, 7, NA)))))))
```

The output of the R-Script (from Console window) is given as follows:

```
> library(data.table)
> library(foreach)
> library(readr)
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:data.table’:

    between, first, last

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> setwd("F:/ACADGILD - Online Course/1. DATA
SETS/BlogFeedback")
> getwd()
[1] "F:/ACADGILD - Online Course/1. DATA
SETS/BlogFeedback"
>
> blogData_train <- read_csv("F:/ACADGILD - Online Course/1. DATA
SETS/BlogFeedback/blogData_train.csv")
Parsed with column specification:
cols(
  .default = col_double()
)
See spec(...) for full column specifications.
|================================================================================
=====================================| 100%   62 MB
> # retrieve filenames of test sets
> test_filenames = list.files(pattern = "blogData_test")
>
> # load and combine dataset
> train = fread("blogData_train.csv")
> fbtest = foreach(i = 1:length(test_filenames), .combine = rbind) %do% {
+ temp = fread(test_filenames[i], header = F)
+ }
>
> # Assign variable names to the train and test data set
> colnames(blogData_train) <-
c("plikes","checkin","talking","category","d5","d6","d7","d8","d9","d10","d11
","d12",
+
"d13","d14","d15","d16","d17","d18","d19","d20","d21","d22","d23","d24","d25"
,"d26",
+
"d27","d28","d29","cc1","cc2","cc3","cc4","cc5","basetime","postlength","post
shre",
+
"postpromo","Hhrs","sun","mon","tue","wed","thu","fri","sat","basesun","basem
on",
+
```

```
"basetue","basewed","basethu","basefri","basesat","target")
> dim(blogData_train)
[1] 52396   281
> dim(fbtest)
[1] 7624 281

> View(blogData_train)
```

Untitled1* | blogData_train | Untitled4* | Untitled3* | fbtest

Filter

	plikes	checkin	talking	category	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18
1	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
2	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
3	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
4	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
5	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
6	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
7	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
8	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
9	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
10	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
11	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
12	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
13	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
14	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
15	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
16	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
17	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	
18	40.30467	53.84566	0	401	15	15.52416	32.44188	0	377	3	14.04423	32.61542	0	377	2	34.56757	48.47518	

Showing 1 to 20 of 52,396 entries

```
> View(fbtest)
```

Untitled1* | blogData_train | Untitled4* | Untitled3* | fbtest

Filter

	V1	V145	V144	V2	V3	V142	V143	V4	V5	V146	V147	V6	V7	V148	V149	V8	V9	V1
1	10.63066000	0	0	17.8829920	1	0	0	259	5.0	0	0	4.01827600	10.3967900	0	0	0	235	
2	43.43582500	0	0	75.5904850	0	0	0	634	20.0	0	0	15.99858950	44.5608700	0	0	0	473	
3	1.73333330	0	0	3.0433900	0	0	1	9	0.0	0	0	0.73333335	1.5260698	0	0	0	5	
4	27.23021500	0	0	45.9709500	0	0	1	371	14.0	0	0	10.78417300	24.2099420	0	0	0	228	
5	4.50000000	0	0	6.6770754	0	0	1	18	0.5	0	0	3.00000000	4.0000000	0	0	0	10	
6	156.40298000	0	0	246.0559800	0	0	1	970	28.0	0	1	76.14925400	131.9008300	0	0	0	725	
7	10.50931600	0	0	36.5939830	0	0	1	191	1.0	0	0	3.60248450	20.6338310	0	0	0	179	
8	123.86919000	0	0	129.5662200	0	0	1	1065	87.0	0	0	43.32897000	62.7741470	0	0	0	491	
9	22.46341500	0	0	42.1849000	0	0	0	188	7.5	0	0	8.21951200	25.0204930	0	0	0	174	
10	0.00000000	0	0	0.0000000	0	0	1	0	0.0	0	0	0.00000000	0.0000000	0	0	0	0	
11	0.15550756	0	0	0.6683261	0	0	0	7	0.0	0	0	0.07559396	0.4113776	0	0	0	5	
12	16.59357500	0	0	19.6713640	1	0	0	144	10.0	0	0	6.51244970	11.0512150	0	0	0	111	
13	0.37869823	0	0	1.0817565	0	0	1	4	0.0	0	0	0.03550296	0.2146551	0	0	0	2	
14	49.44236800	0	0	112.6201250	1	0	0	849	9.0	0	0	20.44548200	62.6193900	0	0	0	506	
15	122.81293000	0	0	109.9611000	0	0	1	1069	89.0	0	0	44.89454300	74.5475300	0	0	0	1046	
16	56.51209300	0	0	77.4428300	0	0	1	438	32.0	0	0	19.29653000	49.2213440	0	0	0	432	
17	43.43582500	0	0	75.5904850	0	0	1	634	20.0	0	0	15.99858950	44.5608700	0	0	0	473	
18	10.63066000	0	0	17.8829920	1	0	0	259	5.0	0	0	4.01827600	10.3967900	0	0	0	235	

```
> str(blogData_train)
Classes ‘spec_tbl_df’, ‘tbl_df’, ‘tbl’ and 'data.frame':    52396 obs. of
281 variables:
 $ plikes    : num  40.3  40.3 40.3  40.3  40.3 ...
 $ checkin   : num  53.8  53.8 53.8  53.8  53.8 ...
 $ talking   : num  0  0 0  0 0 0  0  0 0 ...
 $ category  : num  401 401 401  401  401 401 401 401 401 401 ...
 $ d5        : num  15 15  15 15 15  15 15 15 15 15 ...
 $ d6        : num  15.5  15.5 15.5  15.5  15.5 ...
 $ d7        : num  32.4  32.4 32.4  32.4  32.4 ...
 $ d8        : num  0  0 0  0 0 0  0  0 0 ...
 $ d9        : num  377 377 377  377  377 377 377 377 377 377 ...
 $ d10       : num  3  3 3  3 3 3  3  3 3 ...
 $ d11       : num  14 14  14 14 14 ...
 $ d12       : num  32.6  32.6 32.6  32.6  32.6 ...
 $ d13       : num  0  0 0  0 0 0  0  0 0 ...
 $ d14       : num  377 377 377  377  377 377 377 377 377 377 ...
 $ d15       : num  2  2 2  2 2 2  2  2 2 ...
 $ d16       : num  34.6  34.6 34.6  34.6  34.6 ...
 $ d17       : num  48.5  48.5 48.5  48.5  48.5 ...
 $ d18       : num  0  0 0  0 0 0  0  0 0 ...
 $ d19       : num  378 378 378  378  378 378 378 378 378 378 ...
 $ d20       : num  12 12  12 12 12  12 12 12 12 12 ...
 $ d21       : num  1.48  1.48 1.48  1.48  1.48 ...
 $ d22       : num  46.2  46.2 46.2  46.2  46.2 ...
 $ d23       : num  -356 -356 -356  -356 -356 -356 -356 -356 -356 -356 ...
 $ d24       : num  377 377 377  377  377 377 377 377 377 377 ...
 $ d25       : num  0  0 0  0 0 0  0  0 0 ...
 $ d26       : num  1.08  1.08 1.08  1.08  1.08 ...
 $ d27       : num  1.8 1.8 1.8  1.8  1.8 ...
 $ d28       : num  0  0 0  0 0 0  0  0 0 ...
 $ d29       : num  11 11  11 11 11  11 11 11 11 11 ...
 $ cc1       : num  0  0 0  0 0 0  0  0 0 ...
 $ cc2       : num  0.4 0.4 0.4  0.4  0.4 ...
 $ cc3       : num  1.08  1.08 1.08  1.08  1.08 ...
 $ cc4       : num  0  0 0  0 0 0  0  0 0 ...
 $ cc5       : num  9  9 9  9 9 9  9  9 9 ...
 $ basetime  : num  0  0 0  0 0 0  0  0 0 ...
 $ postlength: num  0.378  0.378 0.378  0.378 0.378 ...
 $ postshre  : num  1.07  1.07 1.07  1.07  1.07 ...
 $ postpromo : num  0  0 0  0 0 0 0  0  0 0 ...
 $ Hhrs      : num  9  9 9  9 9 9 9  9  9 9 ...
 $ sun       : num  0  0 0  0 0 0 0  0  0 0 ...
 $ mon       : num  0.973  0.973 0.973  0.973  0.973 ...
 $ tue       : num  1.7 1.7  1.7 1.7  1.7 ...
 $ wed       : num  0  0 0  0 0 0 0  0  0 0 ...
 $ thu       : num  10 10  10 10 10 10  10 10  10 10 ...
 $ fri       : num  0  0 0  0 0 0 0  0  0 0 ...
 $ sat       : num  0.0229 0.0229 0.0229 0.0229 0.0229 ...
 $ basesun   : num  1.52 1.52  1.52 1.52 1.52 ...
 $ basemon   : num  -8 -8  -8 -8 -8 -8  -8 -8  -8 -8 ...
 $ basetue   : num  9  9 9  9 9 9 9  9  9 9 ...
 $ basewed   : num  0  0 0  0 0 0 0  0  0 0 ...
 $ basethu   : num  6  6 2  3 6 6 3 30 30 0 ...
 $ basefri   : num  2  2 2 1 0  0 1 27 27 0 ...
 $ basesat   : num  4  4 0  2 2 2 2 1  1 0 ...
 $ target    : num  5  5 2  2 5 5 2 2  2 0 ...
 $ NA        : num  -2 -2 2 -1 -2 -2 -1 26 26 0 ...
 $ NA        : num  0  0 0  0 0 0 0 0  0 2 ...
 $ NA        : num  0  0 0  0 0 0 0 0  0 2 ...
 $ NA        : num  0  0 0  0 0 0 0 0  0 0 ...
 $ NA        : num  0  0 0  0 0 0 0 0  0 2 ...
 $ NA        : num  0  0 0  0 0 0 0 0  0 2 ...
```

```
 $ NA        : num  35 35  10 34 59 59  34 58  58 11 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0 0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0 0 0  0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0 0 0  0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
 $ NA        : num  0  0 0  0  0 0 0 0  0  0 ...
$ NA         : num   0  0  0  0  0  0  0  0  0 0 ...
$ NA         : num   0  0  0  0  0  0  0  0  0 0 ...
  [list output truncated]
- attr(*, "spec")=
  .. cols(
  ..   `40.30467` = col_double(),
  ..   `53.845657` = col_double(),
  ..   `0.0` = col_double(),
  ..   `401.0` = col_double(),
  ..   `15.0` = col_double(),
  ..   `15.52416` = col_double(),
  ..   `32.44188` = col_double(),
  ..   `0.0_1` = col_double(),
  ..   `377.0` = col_double(),
  ..   `3.0` = col_double(),
  ..   `14.044226` = col_double(),
  ..   `32.615417` = col_double(),
  ..   `0.0_2` = col_double(),
  ..   `377.0_1` = col_double(),
  ..   `2.0` = col_double(),
  ..   `34.567566` = col_double(),
  ..   `48.475178` = col_double(),
  ..   `0.0_3` = col_double(),
  ..   `378.0` = col_double(),
  ..   `12.0` = col_double(),
  ..   `1.4799345` = col_double(),
```

```
..   `46.18691` = col_double(),
..   `-356.0` = col_double(),
..   `377.0_2` = col_double(),
..   `0.0_4` = col_double(),
..   `1.0761671` = col_double(),
..   `1.795416` = col_double(),
..   `0.0_5` = col_double(),
..   `11.0` = col_double(),
..   `0.0_6` = col_double(),
..   `0.4004914` = col_double(),
..   `1.0780969` = col_double(),
..   `0.0_7` = col_double(),
..   `9.0` = col_double(),
..   `0.0_8` = col_double(),
..   `0.37755936` = col_double(),
..   `1.07421` = col_double(),
..   `0.0_9` = col_double(),
..   `9.0_1` = col_double(),
..   `0.0_10` = col_double(),
..   `0.972973` = col_double(),
..   `1.704671` = col_double(),
..   `0.0_11` = col_double(),
..   `10.0` = col_double(),
..   `0.0_12` = col_double(),
..   `0.022932023` = col_double(),
..   `1.521174` = col_double(),
..   `-8.0` = col_double(),
..   `9.0_2` = col_double(),
..   `0.0_13` = col_double(),
..   `2.0_1` = col_double(),
..   `2.0_2` = col_double(),
..   `0.0_14` = col_double(),
..   `2.0_3` = col_double(),
..   `2.0_4` = col_double(),
..   `0.0_15` = col_double(),
..   `0.0_16` = col_double(),
..   `0.0_17` = col_double(),
..   `0.0_18` = col_double(),
..   `0.0_19` = col_double(),
..   `10.0_1` = col_double(),
..   `0.0_20` = col_double(),
..   `0.0_21` = col_double(),
..   `0.0_22` = col_double(),
..   `0.0_23` = col_double(),
..   `0.0_24` = col_double(),
..   `0.0_25` = col_double(),
..   `0.0_26` = col_double(),
..   `0.0_27` = col_double(),
..   `0.0_28` = col_double(),
..   `0.0_29` = col_double(),
..   `0.0_30` = col_double(),
..   `0.0_31` = col_double(),
..   `0.0_32` = col_double(),
..   `0.0_33` = col_double(),
..   `0.0_34` = col_double(),
..   `0.0_35` = col_double(),
..   `0.0_36` = col_double(),
..   `0.0_37` = col_double(),
..   `0.0_38` = col_double(),
..   `0.0_39` = col_double(),
..   `0.0_40` = col_double(),
..   `0.0_41` = col_double(),
..   `0.0_42` = col_double(),
```

```
..   `0.0_43` = col_double(),
..   `0.0_44` = col_double(),
..   `0.0_45` = col_double(),
..   `0.0_46` = col_double(),
..   `0.0_47` = col_double(),
..   `0.0_48` = col_double(),
..   `0.0_49` = col_double(),
..   `0.0_50` = col_double(),
..   `0.0_51` = col_double(),
..   `0.0_52` = col_double(),
..   `0.0_53` = col_double(),
..   `0.0_54` = col_double(),
..   `0.0_55` = col_double(),
..   `0.0_56` = col_double(),
..   `0.0_57` = col_double(),
..   `0.0_58` = col_double(),
..   `0.0_59` = col_double(),
..   `0.0_60` = col_double(),
..   `0.0_61` = col_double(),
..   `0.0_62` = col_double(),
..   `0.0_63` = col_double(),
..   `0.0_64` = col_double(),
..   `0.0_65` = col_double(),
..   `0.0_66` = col_double(),
..   `0.0_67` = col_double(),
..   `0.0_68` = col_double(),
..   `0.0_69` = col_double(),
..   `0.0_70` = col_double(),
..   `0.0_71` = col_double(),
..   `0.0_72` = col_double(),
..   `0.0_73` = col_double(),
..   `0.0_74` = col_double(),
..   `0.0_75` = col_double(),
..   `0.0_76` = col_double(),
..   `0.0_77` = col_double(),
..   `0.0_78` = col_double(),
..   `0.0_79` = col_double(),
..   `0.0_80` = col_double(),
..   `0.0_81` = col_double(),
..   `0.0_82` = col_double(),
..   `0.0_83` = col_double(),
..   `0.0_84` = col_double(),
..   `0.0_85` = col_double(),
..   `0.0_86` = col_double(),
..   `0.0_87` = col_double(),
..   `0.0_88` = col_double(),
..   `0.0_89` = col_double(),
..   `0.0_90` = col_double(),
..   `0.0_91` = col_double(),
..   `0.0_92` = col_double(),
..   `0.0_93` = col_double(),
..   `0.0_94` = col_double(),
..   `0.0_95` = col_double(),
..   `0.0_96` = col_double(),
..   `0.0_97` = col_double(),
..   `0.0_98` = col_double(),
..   `0.0_99` = col_double(),
..   `0.0_100` = col_double(),
..   `0.0_101` = col_double(),
..   `0.0_102` = col_double(),
..   `0.0_103` = col_double(),
..   `0.0_104` = col_double(),
..   `0.0_105` = col_double(),
```

```
..   `0.0_106` = col_double(),
..   `0.0_107` = col_double(),
..   `0.0_108` = col_double(),
..   `0.0_109` = col_double(),
..   `0.0_110` = col_double(),
..   `0.0_111` = col_double(),
..   `0.0_112` = col_double(),
..   `0.0_113` = col_double(),
..   `0.0_114` = col_double(),
..   `0.0_115` = col_double(),
..   `0.0_116` = col_double(),
..   `0.0_117` = col_double(),
..   `0.0_118` = col_double(),
..   `0.0_119` = col_double(),
..   `0.0_120` = col_double(),
..   `0.0_121` = col_double(),
..   `0.0_122` = col_double(),
..   `0.0_123` = col_double(),
..   `0.0_124` = col_double(),
..   `0.0_125` = col_double(),
..   `0.0_126` = col_double(),
..   `0.0_127` = col_double(),
..   `0.0_128` = col_double(),
..   `0.0_129` = col_double(),
..   `0.0_130` = col_double(),
..   `0.0_131` = col_double(),
..   `0.0_132` = col_double(),
..   `0.0_133` = col_double(),
..   `0.0_134` = col_double(),
..   `0.0_135` = col_double(),
..   `0.0_136` = col_double(),
..   `0.0_137` = col_double(),
..   `0.0_138` = col_double(),
..   `0.0_139` = col_double(),
..   `0.0_140` = col_double(),
..   `0.0_141` = col_double(),
..   `0.0_142` = col_double(),
..   `0.0_143` = col_double(),
..   `0.0_144` = col_double(),
..   `0.0_145` = col_double(),
..   `0.0_146` = col_double(),
..   `0.0_147` = col_double(),
..   `0.0_148` = col_double(),
..   `0.0_149` = col_double(),
..   `0.0_150` = col_double(),
..   `0.0_151` = col_double(),
..   `0.0_152` = col_double(),
..   `0.0_153` = col_double(),
..   `0.0_154` = col_double(),
..   `0.0_155` = col_double(),
..   `0.0_156` = col_double(),
..   `0.0_157` = col_double(),
..   `0.0_158` = col_double(),
..   `0.0_159` = col_double(),
..   `0.0_160` = col_double(),
..   `0.0_161` = col_double(),
..   `0.0_162` = col_double(),
..   `0.0_163` = col_double(),
..   `0.0_164` = col_double(),
..   `0.0_165` = col_double(),
..   `0.0_166` = col_double(),
..   `0.0_167` = col_double(),
..   `0.0_168` = col_double(),
```

```
..   `0.0_169` = col_double(),
..   `0.0_170` = col_double(),
..   `0.0_171` = col_double(),
..   `0.0_172` = col_double(),
..   `0.0_173` = col_double(),
..   `0.0_174` = col_double(),
..   `0.0_175` = col_double(),
..   `0.0_176` = col_double(),
..   `0.0_177` = col_double(),
..   `0.0_178` = col_double(),
..   `0.0_179` = col_double(),
..   `0.0_180` = col_double(),
..   `0.0_181` = col_double(),
..   `0.0_182` = col_double(),
..   `0.0_183` = col_double(),
..   `0.0_184` = col_double(),
..   `0.0_185` = col_double(),
..   `0.0_186` = col_double(),
..   `0.0_187` = col_double(),
..   `0.0_188` = col_double(),
..   `0.0_189` = col_double(),
..   `0.0_190` = col_double(),
..   `0.0_191` = col_double(),
..   `0.0_192` = col_double(),
..   `0.0_193` = col_double(),
..   `0.0_194` = col_double(),
..   `0.0_195` = col_double(),
..   `0.0_196` = col_double(),
..   `0.0_197` = col_double(),
..   `0.0_198` = col_double(),
..   `0.0_199` = col_double(),
..   `0.0_200` = col_double(),
..   `0.0_201` = col_double(),
..   `0.0_202` = col_double(),
..   `0.0_203` = col_double(),
..   `0.0_204` = col_double(),
..   `0.0_205` = col_double(),
..   `0.0_206` = col_double(),
..   `0.0_207` = col_double(),
..   `0.0_208` = col_double(),
..   `0.0_209` = col_double(),
..   `0.0_210` = col_double(),
..   `0.0_211` = col_double(),
..   `0.0_212` = col_double(),
..   `0.0_213` = col_double(),
..   `0.0_214` = col_double(),
..   `0.0_215` = col_double(),
..   `0.0_216` = col_double(),
..   `0.0_217` = col_double(),
..   `0.0_218` = col_double(),
..   `0.0_219` = col_double(),
..   `0.0_220` = col_double(),
..   `0.0_221` = col_double(),
..   `0.0_222` = col_double(),
..   `0.0_223` = col_double(),
..   `0.0_224` = col_double(),
..   `1.0` = col_double(),
..   `0.0_225` = col_double(),
..   `0.0_226` = col_double(),
..   `0.0_227` = col_double(),
..   `0.0_228` = col_double(),
..   `0.0_229` = col_double(),
..   `1.0_1` = col_double(),
```

```
  ..   `0.0_230` = col_double(),
  ..   `0.0_231` = col_double(),
  ..   `0.0_232` = col_double(),
  ..   `0.0_233` = col_double(),
  ..   `0.0_234` = col_double(),
  ..   `0.0_235` = col_double(),
  ..   `0.0_236` = col_double(),
  ..   `1.0_2` = col_double()
  .. )
```

```
> str(fbtest)
```

```
Classes ‘data.table’ and 'data.frame':7624 obs. of  281 variables:
 $ V1  : num  10.63 43.44 1.73 27.23 4.5 ...
 $ V145: num  0 0 0  0 0 0 0 0 0  0 ...
 $ V144: num  0 0 0  0 0 0 0 0 0  0 ...
 $ V2  : num  17.88 75.59 3.04 45.97 6.68 ...
 $ V3  : num  1 0 0  0 0 0 0 0 0  0 ...
 $ V142: num  0 0 0  0 0 0 0 0 0  0 ...
 $ V143: num  0 0 1  1 1 1 1 1 0  1 ...
 $ V4  : num  259 634 9 371 18 ...
 $ V5  : num  5 20 0 14 0.5 28 1 87 7.5 0 ...
 $ V146: num  0 0 0  0 0 0 0 0 0  0 ...
 $ V147: num  0 0 0  0 0 1 0 0 0  0 ...
 $ V6  : num 4.018 15.999 0.733 10.784 3 ...
 $ V7  : num 10.4 44.56 1.53 24.21 4 ...
 $ V148: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V149: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V8  : num 0 0 0 0 0 0 0 0 0 0 ...
 $ V9  : num 235 473 5 228 10 725 179 491 174 0 ...
 $ V150: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V151: num 0 1 1 0 0 1 1 0 0 1 ...
 $ V10 : num 1 2 0 4 0.5 16 0 19.5 1.5 0 ...
 $ V11 : num 3.817 15.47 0.667 9.998 1.333 ...
 $ V152: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V153: num 0 0 1 0 0 1 0 0 0 0 ...
 $ V12 : num 10.3 44.69 1.53 24.4 2.56 ...
 $ V13 : num 0 0 0 0 0 0 0 0 0 0 ...
 $ V154: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V155: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V14 : num 235 473 5 228 7 725 179 491 174 0 ...
 $ V15 : num 1 1 0 2 0 3 0 14 1 0 ...
 $ V156: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V157: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V16 : num 9.78 40.97 1.13 22.56 2.83 ...
 $ V17 : num 16.07 70.31 1.82 39.76 3.67 ...
 $ V158: num 0 0 1 1 0 1 1 0 0 1 ...
 $ V159: num 0 0 1 0 0 1 0 0 0 0 ...
 $ V18 : num 1 0 0 0 0 0 0 0 0 0 ...
 $ V19 : num 192 479 5 337 8 913 189 786 186 0 ...
 $ V160: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V161: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V20 : num 5 18 0 10 0.5 26 0 74 5.5 0 ...
 $ V21 : num 0.201 0.5289 0.0667 0.7866 1.6667 ...
 $ V162: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V163: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V22 : num 13.95 62.13 1.73 30.36 2.21 ...
 $ V23 : num -229 -461 -5 -156 0 -519 -178 -418 -161 0 ...
 $ V164: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V165: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V24 : num 217 473 4 228 6 725 170 491 174 0 ...
 $ V25 : num 0 0 0 0 0.5 2 0 -3 0 0 ...
 $ V166: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V167: num 0 0 0 0 0 0 0 0 0 0 ...
```

```
 $ V26  : num 0.252 0.193 0.333 0.11 0 ...
 $ V27  : num 0.904 0.458 0.699 0.356 0 ...
 $  V168: num 0 0 0 0 0 0 0 0 0 0 ...
 $  V169: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V28  : num 0 0 0 0 0 0 0 0 0 0 ...
 $ V29  : num 14 2 2 2 0 0 6 0 1 0 ...
 $  V170: num 0 0 1 0 0 1 0 0 0 0 ...
 $  V171: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V30  : num 0 0 0 0 0 0 0 0 0 0 ...
 $ V31  : num 0.0944 0.0733 0.1333 0.0432 0 ...
 $  V172: num 0 0 0 0 0 0 0 0 0 0 ...
 $  V173: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V32  : num 0.507 0.286 0.34 0.215 0 ...
 $ V33  : num 0 0 0 0 0 0 0 0 0 0 ...
 $  V174: num 0 0 0 0 0 0 0 0 1 0 ...
 $  V175: num 0 0 0 0 0 0 0 0 0 0 ...
 $ V34  : num 12 2 1 2 0 0 5 0 1 0 ...
 $ V35  : num 0 0 0 0 0 0 0 0 0 0 ...
 $  V176: num 0 0 0 0 0 0 0 0 0 0 ...
 $  V177: num 0 0 0 0 0 0 0 0 0 0 ...
$ V36 : num  0.0919 0.0677 0.1333 0.0408 0 ...
$ V37 : num  0.504 0.278 0.34 0.21 0 ...
$ V178: num  0  0 0 0  0 0  0 0 0 0 ...
$ V179: num  0  0 0 0  0 0  0 0 0 0 ...
$ V38 : num  0  0 0 0  0 0  0 0 0 0 ...
$ V39 : num  12 2 1 2 0 0 5 0 1 0 ...
$ V180: num  0  0 1 0  0 1  1 0 0 0 ...
$ V181: num  0  0 1 0  0 0  0 0 0 0 ...
$ V40 : num  0  0 0 0  0 0  0 0 0 0 ...
$ V41 : num  0.2335 0.1763  0.2 0.0983 0 ...
$ V182: num  0  0 0 0  0 0  0 0 0 0 ...
$ V183: num  0  0 0 0  0 1  0 0 0 0 ...
$ V42 : num  0.855 0.43 0.4 0.321 0 ...
$ V43 : num  0  0 0 0  0 0 0  0 0 0 ...
$ V184: num  0  0 0 0  0 0  0 0 0 0 ...
$ V185: num  0  0 0 0  0 0  0 0 0 0 ...
$ V44 : num  13 2 1 2 0 0 5 0 1 0 ...
$ V45 : num  0  0 0 0  0 0  0 0 0 0 ...
$ V186: num  0  0 0 0  0 0  0 0 0 0 ...
$ V187: num  0  0 0 0  0 0  0 0 0 0 ...
$ V46 : num  0.00245  0.00564 0 0.0024 0 ...
$ V47 : num  0.675 0.404  0.365 0.29 0 ...
$ V188: num  0  0 0 0  0 0  0 0 0 0 ...
$ V189: num  0  0 0 0  0 0  0 0 0 0 ...
$ V48 : num  -10 -2 -1 -2 0 0 -5 0 -1 0 ...
$ V49 : num  12 2 1 2 0 0 5 0 1 0 ...
$ V190: num  0  0 0 0  0 0  0 0 0 0 ...
$ V191: num  0  0 1 0  0 1  1 0 0 1 ...
 [list output truncated]
 - attr(*, ".internal.selfref")=<externalptr>
>
> train <- blogData_train; test <- fbtest
```

```
> head(train); head(test)
# A tibble: 6 x 281
  plikes checkin talking category    d5    d6    d7    d8    d9   d10   d11
d12   d13  d14   d15  d16   d17   d18   d19   d20
    <dbl>   <dbl>   <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl> <dbl>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
1    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
2    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
3    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
4    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
5    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
6    40.3    53.8       0      401     15  15.5  32.4     0   377     3 14.0
32.6      0   377     2 34.6  48.5     0   378    12
# ... with 261 more variables: d21 <dbl>, d22 <dbl>, d23 <dbl>, d24 <dbl>,
d25 <dbl>, d26 <dbl>, d27 <dbl>, d28 <dbl>,
#   d29 <dbl>, cc1 <dbl>, cc2 <dbl>, cc3 <dbl>, cc4 <dbl>, cc5 <dbl>,
basetime <dbl>, postlength <dbl>, postshre <dbl>,
#   postpromo <dbl>, Hhrs <dbl>, sun <dbl>, mon <dbl>, tue <dbl>, wed <dbl>,
thu <dbl>, fri <dbl>, sat <dbl>, basesun <dbl>,
#   basemon <dbl>, basetue <dbl>, basewed <dbl>, basethu <dbl>, basefri
<dbl>, basesat <dbl>, target <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,

#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, ...
            V1 V145  V144         V2 V3 V142  V143 V4   V5 V146 V147         V6
V7 V148 V149 V8  V9  V150 V151  V10
1:   10.630660    0     0 17.882992  1    0     0 259   5.0    0    0  4.0182760
10.39679    0    0  0 235     0     0  1.0
2:   43.435825    0     0 75.590485  0    0     0 634  20.0    0    0 15.9985895
44.56087    0    0  0 473     0     1  2.0
3:    1.733333    0     0   3.043390  0    0     1   9   0.0    0    0  0.7333333
1.52607    0     0  0  5      0    1   0.0
           V11 V152  V153        V12 V13 V154 V155 V14 V15 V156 V157         V16
V17  V158 V159 V18  V19  V160 V161   V20
1:   3.8172395    0     0 10.297346    0     0    0 235    1     0     0   9.776869
16.073494    0     0     1  192     0    0  5.0
2: 15.4696760    0     0 44.685085    0     0    0 473    1     0     0  40.971790
70.307840    0     0     0  479     0    0 18.0
3:   0.6666667   0     1   1.534782    0     0    0    5    0     0     0   1.133333
1.820867     1    1     0   5     0      0  0.0
           V21  V162  V163          V22   V23   V164  V165  V24  V25  V166  V167
V26          V27 V168   V169 V28  V29   V170 V171  V30
1:  0.20103656     0     0   13.948867 -229      0     0 217 0.0     0     0
0.2517731 0.9038038    0      0   0   14     0      0   0
2:  0.52891400     0     0   62.134968 -461      0     0 473 0.0     0     0
0.1932299 0.4576994    0      0   0    2     0      0   0
3:  0.06666667     0     0    1.730767   -5      0     0    4 0.0     0     0
0.3333333 0.6992059    0      0   0    2     1      0   0
           V31 V172   V173        V32  V33  V174   V175 V34   V35  V176  V177         V36
```

```
v37  v178 v179v38  v39  v180 v181 v40
1: 0.09438080    0     0 0.5067316   0    0     0 12    0    0    0  0.09192581
0.5042160    0    0    0  12    0   0   0
2: 0.07334273    0     0 0.2864750   0    0     0  2    0    0    0  0.06770099
0.2778884    0    0    0   2    0   0   0
3: 0.13333334    0     0 0.3399347   0    0     0  1    0    0    0  0.13333334
0.3399347    0    0    0   1    1   1   0
          v41 v182  v183       v42  v43 v184  v185 v44  v45  v186  v187         v46
v47  v188 v189v48  v49  v190 v191 v50  v51  v192
1: 0.23349700    0     0 0.8547111   0    0     0 13    0    0    0 0.002454992
0.6747285    0    0  -10  12    0   0   0   35   0
2: 0.17630465    0     0 0.4297832   0    0     0  2    0    0    0 0.005641749
0.4044489    0    0   -2   2    0   0   0   21   0
3: 0.20000000    0     0 0.4000000   0    0     0  1    0    0    0 0.000000000
0.3651484    0    0   -1   1    0   1   0    2   0
    v193 v52 v53  v194  v195  v54 v55  v196 v197  v56 v57  v198 v199  v58  v59 v200
v201 v60  v61  v202  v203   v62  v63 v204 v205  v64  v65
1:     0  35  0     0     0  35  35    0    0    0  0    0    0    0  0     0
0   0   9    0    0      0   0   0     0  0    0
2:     0   0  2     0     0  21 -2    0    0    0  0    0    0    0  0     0
0   0  62    0    0   696    0   0     0  0    0
3:     0   2  0     0     0   2  2    0    0    2  2    0    0    0  2     0
0   2  13    1    0   8361   0   0     0  1    0
   v206 v207 v66  v67  v208 v209  v68 v69 v210 v211 v70 v71 v212 v213 v72  v73
v214 v215  v74 v75  v216 v217 v76 v77 v218  v219 v78 v79
1:     0    0  0    0     0     0  0   0    0    0   0   0    0    0   0   0
0    0   0   0     0    0    0  0    0     0  0    0
2:     0    0  0    0     0     0  1   0    0    0   0   0    0    0   0   0
0    0   0   0     0    0    0  0    0     0  0    0
3:     0    1  0    1     0     0  1   1    1    0   0   0    0    1   0   0
0    0   0   0     0    0    0  0    0     0  0    1
    v220 v221 v80  v81  v222  v223 v82 v83  v224 v225 v84  v85  v226 v227  v86  v87
v228 v229 v88  v89  v230 v231 v90 v91 v232  v233  v92 v93
1:     0    0  0    0     0     0  0   0    0    0   0   0    0    0   0   0
0    0   0   0     0    0    0  0    0     0  0    0
2:     0    0  0    0     0     0  0   0    0    0   0   0    0    1   0   0
0    0   0   0     0    0    0  0    1     0  0    0
3:     0    0  0    0     0     0  0   0    0    0   0   0    0    1   0   0
1    0   0   0     0    0    0  0    1     0  0    0
    v234 v235 v94  v95  v236  v237 v96 v97  v238 v239 v98  v99  v240 v241 v100  v101
v242 v243 v102 v103 v244 v245 v104 v105 v246 v247
1:     0    0  0    0     0     0  0   0    0    0   0   0    0    0   0     0
0    0     0   0    0    0     0    0   0    0
2:     0    0  0    0     0     0  0   0    0    0   0   0    0    0   0     1
0    0     0   0    0    0     0    0   1    0
3:     0    0  0    0     0     0  1   0    0    0   0   0    0    0   0     1
0    0     1   0    0    0     0    0   1    0
    v106 v107 v248  v249 v108 v109 v250  v251 v110 v111 v252 v253 v112  v113 v254
v255 v114 v115 v256 v257  v116 v117 v258 v259 v118
1:     0    0    0     0    0    0    0     0   0     0   0    0    0     0   0
0    0     0   0    0    0    0     0   0    0
2:     0    0    0     0    0    0    0     0   0     0   0    0    0     0   0
0    0     0   0    0    0    0     0   0    0
3:     0    0    1     0    0    0    0     0   0     0   0    0    0     0   1
0    0     0   0    0    0    0     0   0    0
    v119 v260 v261  v120 v121 v262 v263  v122  v123 v264 v265 v124 v125  v266 v267
v126 v127 v268 v269 v128 v129 v270 v271 v130 v131
1:     0    0    0     0    0    0    0     0   0     0   1    0    0     0   0
0    0     0   0    0    0    0     1   0    0
2:     0    0    0     0    0    0    0     0   0     0   1    0    0     0   0
0    0     0   0    0    0    0     0   0    0
3:     0    0    0     1    0    0    0     1   0     0   1    0    0     0   0
0    0     0   0    0    0    0     1   0    0
```

```
   V272 V273 V132  V133 V274 V275 V134  V135  V276 V277 V136 V137 V278  V279 V138
V139 V280 V281 V140 V141
1:    0    0    0    0    0    0    0     0    0    0    0    0    0     0    0
0    0     4    0    0
2:    0    0    0    0    0    0    0     0    1    0    0    0    0     0    0
1    0     0    0    0
3:    0    0    0    0    0    0    0     0    0    0    0    0    0     0    0
0    0     1    0    0
 [ reached getOption("max.print") -- omitted 3 rows ]
>
> # making the data tidy by constructing single collumn for post publish day
> train$pubday<- ifelse(train$sun ==1, 1, ifelse(train$mon ==1,
2, ifelse(train$tue ==1, 3,
+
ifelse(train$wed ==1, 4, ifelse(train$thu ==1, 5, ifelse(train$fri ==1, 6,
+
ifelse(train$sat ==1, 7, NA)))))))
> # making the data tidy by constructing single collumn for base day
> train$baseday<- ifelse(train$basesun ==1, 1, ifelse(train$basemon ==1, 2,
ifelse(train$basetue ==1, 3,
+
ifelse(train$basewed ==1, 4, ifelse(train$basethu ==1, 5,
+
ifelse(train$basefri ==1, 6, ifelse(train$basesat ==1, 7, NA)))))))
```

Conclusion/Interpretation:

The train and test datasets are read and right features are identified. Now the data set is ready

b. Clean dataset, impute missing values and perform exploratory data analysis.

The R-script for the given problem is as follows:

```
distinct(train) # removing overlapping observations if any
dim(train)
sapply(train, function(x) sum(is.na(x))) # no missing values

correlation <- cor(train,y = NULL, use = "everything",
                   method = c("pearson", "kendall", "spearman"))
corr <- as.data.frame(reshape::melt(correlation))
corr <- corr%>%filter(X1 == "target" & value != 1 & value > 0.32 & value > -
0.32)
corr # good corelations with target variable
library(corrplot)

corrplot.mixed(cor(train[,c(30:32)]))
# Total comments are strongly correlated to correlated with cc3(comments in last 48 to
last 24 hours relative to base date/time)
```

```
df <- train
melt_df <- melt(df)

library(ggplot2)
# Distribution of all the Variables - Histogram
ggplot(melt_df, aes(x=value, fill = variable))+
geom_histogram(bins=10, color = "Blue")+
  facet_wrap(~variable, scales = 'free_x')
df <- log(train[1:39]) par(mfrow=c(1,1))
```

The output of the R-Script (from Console window) is given as follows:

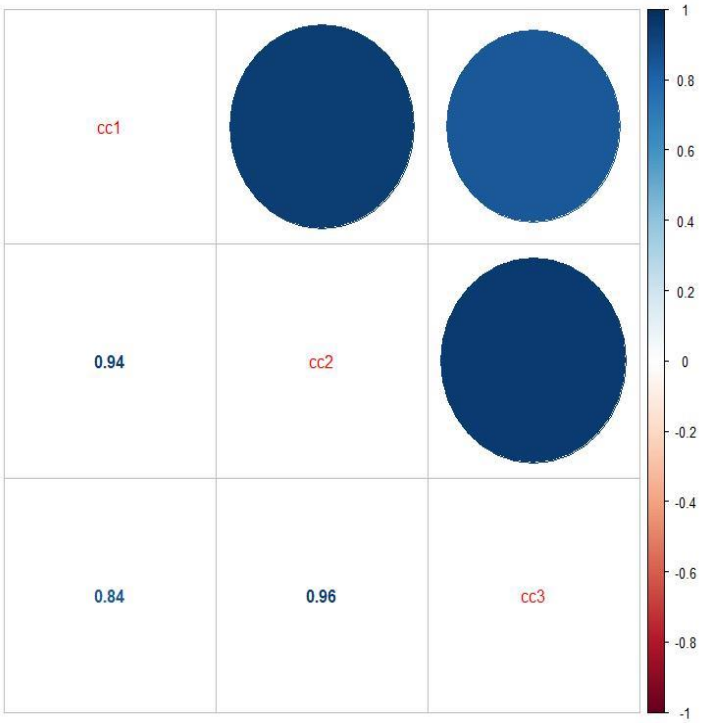
```
> distinct(train)   # removing overlapping observations if any
# A tibble: 49,203 x 283
   plikes checkin talking category    d5    d6    d7    d8    d9   d10   d11
d12    d13  d14   d15   d16   d17   d18   d19
     <dbl>  <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 2   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 3   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 4   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 5   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 6   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 7   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 8   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
 9   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
10   40.3    53.8        0      401    15  15.5  32.4     0   377     3  14.0
32.6      0  377       2  34.6   48.5     0   378
# ... with 49,193 more rows, and 264 more variables: d20 <dbl>, d21 <dbl>,
d22 <dbl>, d23 <dbl>, d24 <dbl>, d25 <dbl>,
#   d26 <dbl>, d27 <dbl>, d28 <dbl>, d29 <dbl>, cc1 <dbl>, cc2 <dbl>, cc3
<dbl>, cc4 <dbl>, cc5 <dbl>, basetime <dbl>,
#   postlength <dbl>, postshre <dbl>, postpromo <dbl>, Hhrs <dbl>, sun <dbl>,
mon <dbl>, tue <dbl>, wed <dbl>, thu <dbl>,
#   fri <dbl>, sat <dbl>, basesun <dbl>, basemon <dbl>, basetue <dbl>,
basewed <dbl>, basethu <dbl>, basefri <dbl>,
#   basesat <dbl>, target <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>,
#   NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA <dbl>, NA
<dbl>, ...
```

```
> dim(train)
[1] 52396   283
> sapply(train, function(x) sum(is.na(x))) # no missing values
    plikes     checkin     talking    category          d5          d6          d7
d8          d9          d10         d11
         0           0           0           0           0           0           0
0           0           0           0
       d12         d13         d14         d15         d16         d17         d18
d19         d20         d21         d22
         0           0           0           0           0           0           0
0           0           0           0
       d23         d24         d25         d26         d27         d28         d29
cc1         cc2         cc3         cc4
         0           0           0           0           0           0           0
0           0           0           0
       cc5    basetime postlength    postshre   postpromo        Hhrs         sun
mon         tue         wed         thu
         0           0           0           0           0           0           0
0           0           0           0

       fri         sat     basesun     basemon     basetue     basewed     basethu
basefri     basesat      target        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
      <NA>        <NA>        <NA>        <NA>        <NA>        <NA>        <NA>
<NA>        <NA>        <NA>        <NA>
         0           0           0           0           0           0           0
0           0           0           0
```

```
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0

          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>
<NA>          <NA>          <NA>          <NA>
             0             0             0             0             0             0             0
0             0             0             0
          <NA>          <NA>          <NA>          <NA>          <NA>          <NA>        pubday
baseday
             0             0             0             0             0             0         41204
34162
>
> correlation <- cor(train,y = NULL, use = "everything",
+                     method = c("pearson", "kendall", "spearman"))
> corr <- as.data.frame(reshape::melt(correlation))
> corr <- corr%>%filter(X1 == "target" & value != 1 & value > 0.32 & value >
-0.32)
> corr  # good corelations with target variable
        X1         X2     value
1   target     plikes 0.7033608
2   target    checkin 0.6582532
3   target   category 0.6140403
4   target         d5 0.6807699
5   target         d6 0.6977038
6   target         d7 0.6697552
7   target         d9 0.5780158
8   target        d10 0.6320845
9   target        d11 0.7018448
10  target        d12 0.6742162
```

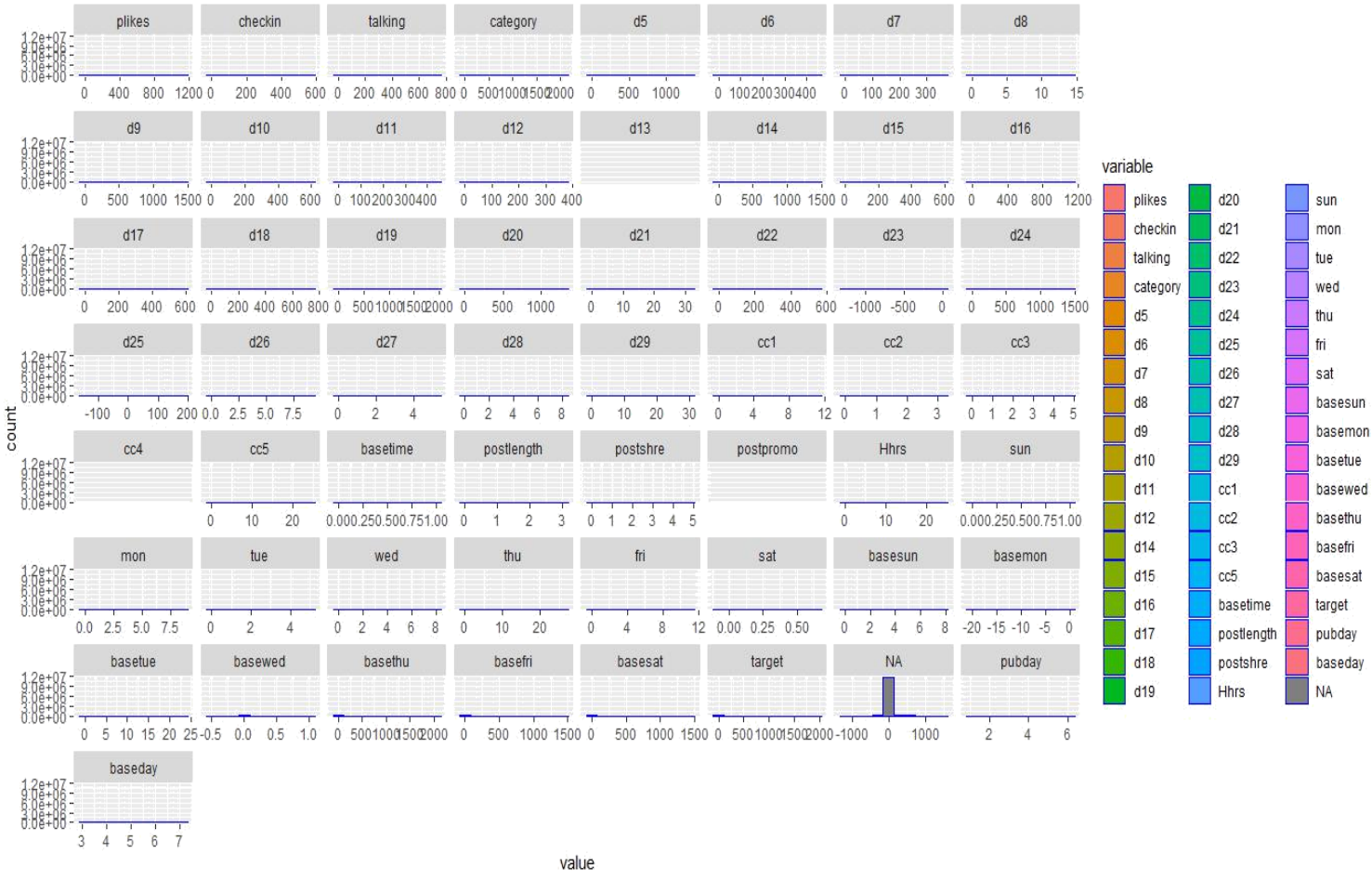
```
11  target        d14 0.5801304
12  target        d15 0.6318017
13  target        d16 0.7053838
14  target        d17 0.6369178
15  target        d19 0.5713231
16  target        d20 0.6814563
17  target        d21 0.5998368
18  target        d22 0.6792232
19  target        d24 0.5784182
20  target        d26 0.4680802
21  target        d27 0.3716850
22  target        d29 0.3436600
23  target        cc1 0.4857482
24  target        cc2 0.4713853
25 target         cc3 0.3958093
 26 target   basetime 0.5353860
 27 target postlength 0.4745144
 28 target   postshre 0.3990222
 29 target        mon 0.4713000
 30 target        tue 0.3742968
 31 target        thu 0.3336524
 32 target        fri 0.4600544
 33 target        sat 0.3211086
 34 target    basesun 0.4087624
 35 target    basethu 0.9755843
 36 target    basefri 0.6832788
 37 target    basesat 0.7092183
 38 target       <NA> 0.5298679
 39 target       <NA> 0.3259848
 40 target       <NA> 0.3617648
 41 target       <NA> 0.5330890
```

```
> library(corrplot)
corrplot 0.84 loaded
> corrplot.mixed(cor(train[,c(30:32)]))
```



```
> df <- train
> melt_df <- melt(df)
> library(ggplot2)
> # Distribution of all the Variables - Histogram
> ggplot(melt_df, aes(x=value, fill = variable))+ +
geom_histogram(bins=10, color = "Blue")+
+   facet_wrap(~variable, scales = 'free_x')
> df <- log(train[1:39])
```

```
> par(mfrow=c(1,1))
```



Conclusion/Interpretation:

- There is a good corelations with target variable
- Total comments are strongly correlated to correlated cc3(comments in last 48 to last 24 hours relative to base date/time)

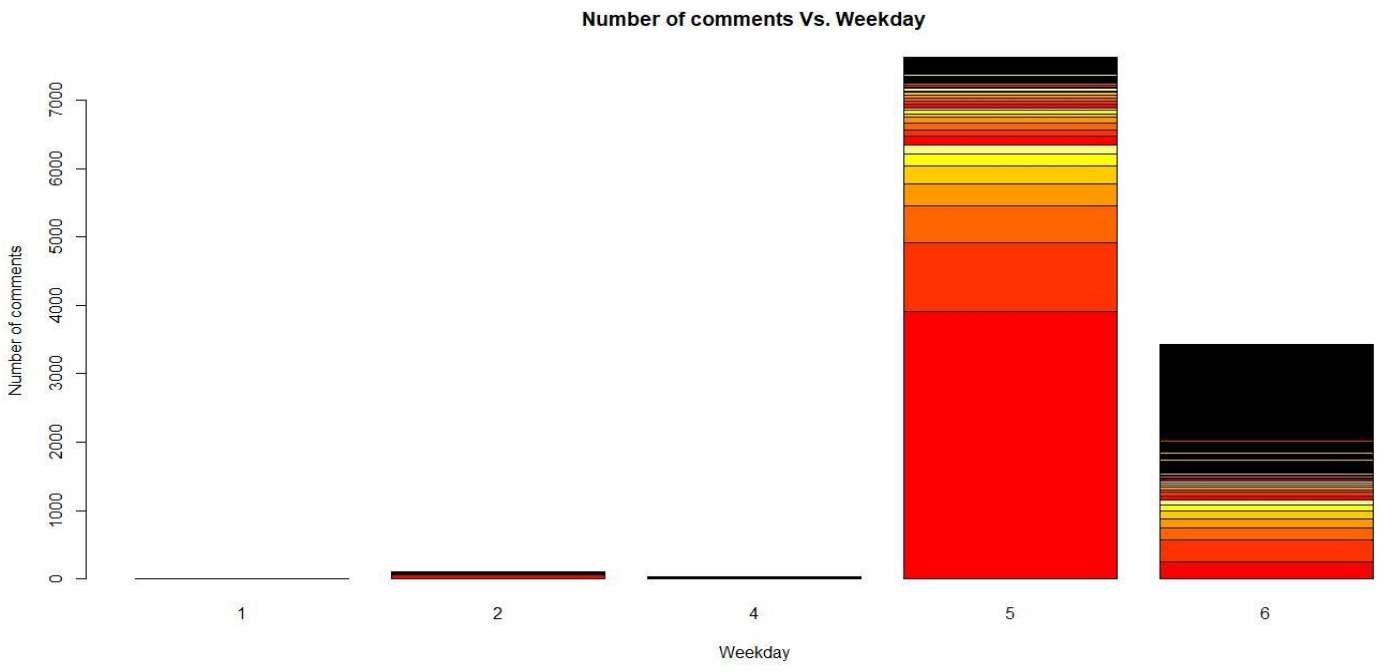
c. Visualize the dataset and make inferences from that.

The R-script for the given problem is as follows:

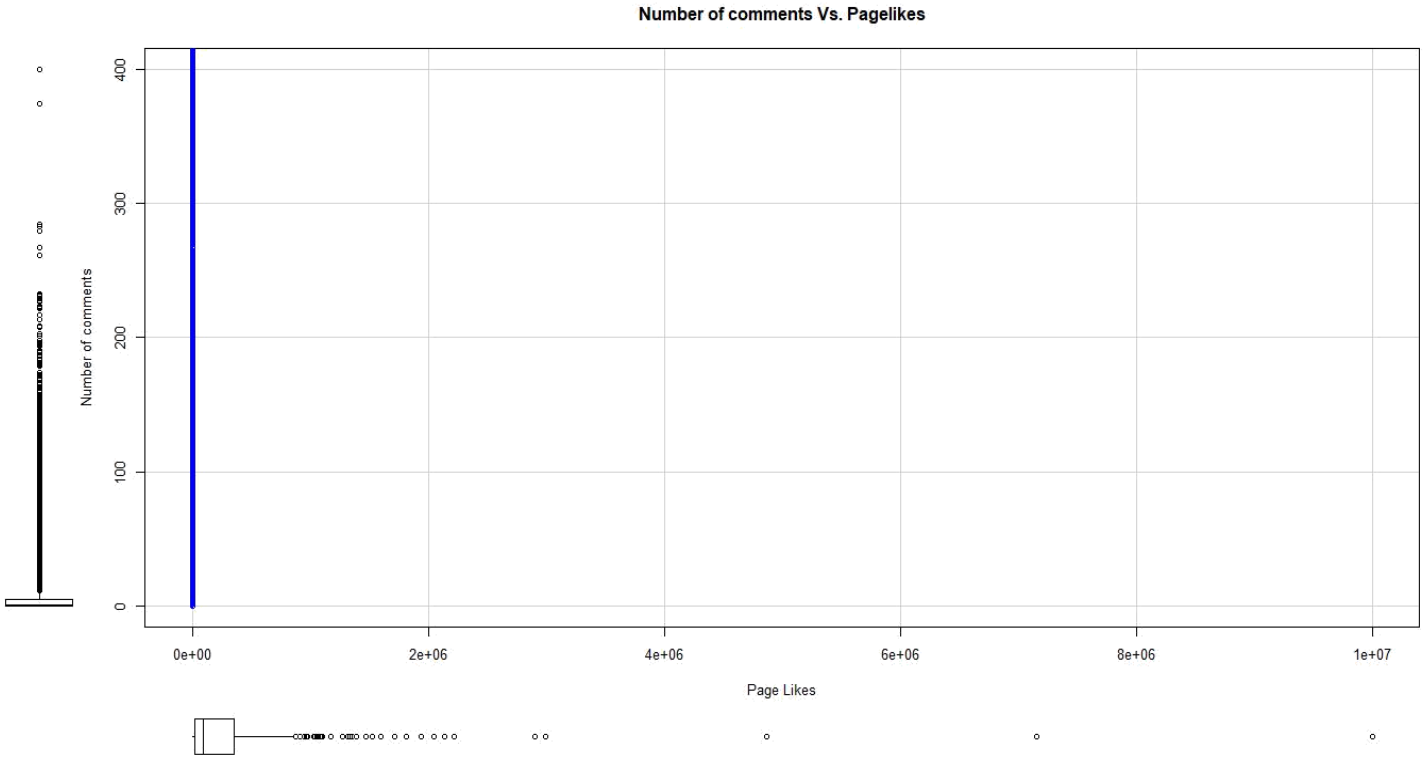
```
barplot(table(train$target, train$pubday), col = heat.colors(7),
      xlab = "Weekday", ylab = "Number of comments", main =
      "Number of comments Vs. Weekday")

library(car)
# number of comments vs Post Likes
scatterplot(train$plikes, train$target , col = "Blue",
        xlab = "Page Likes", ylab = "Number of
        comments", main = "Number of comments Vs.
        Pagelikes", xlim = c(0,10000000), ylim = c(0,400))
abline(lm(plikes~target, data = train), col = "red")

# Number of comments Vs Post length
scatterplot(train$postlength, train$target , col = "Red",
        xlab = "Post Length", ylab = "Number of comments",
        main = "Number of comments Vs. Psot Length",
        ylim = c(0,400), xlim = c(0,5000))
abline(lm(postlength~target, data = train), col= "blue")

hist(train$target, breaks = 1000, xlim = c(0,10) )
```

The output of the R-Script (from Console window) is given as follows:

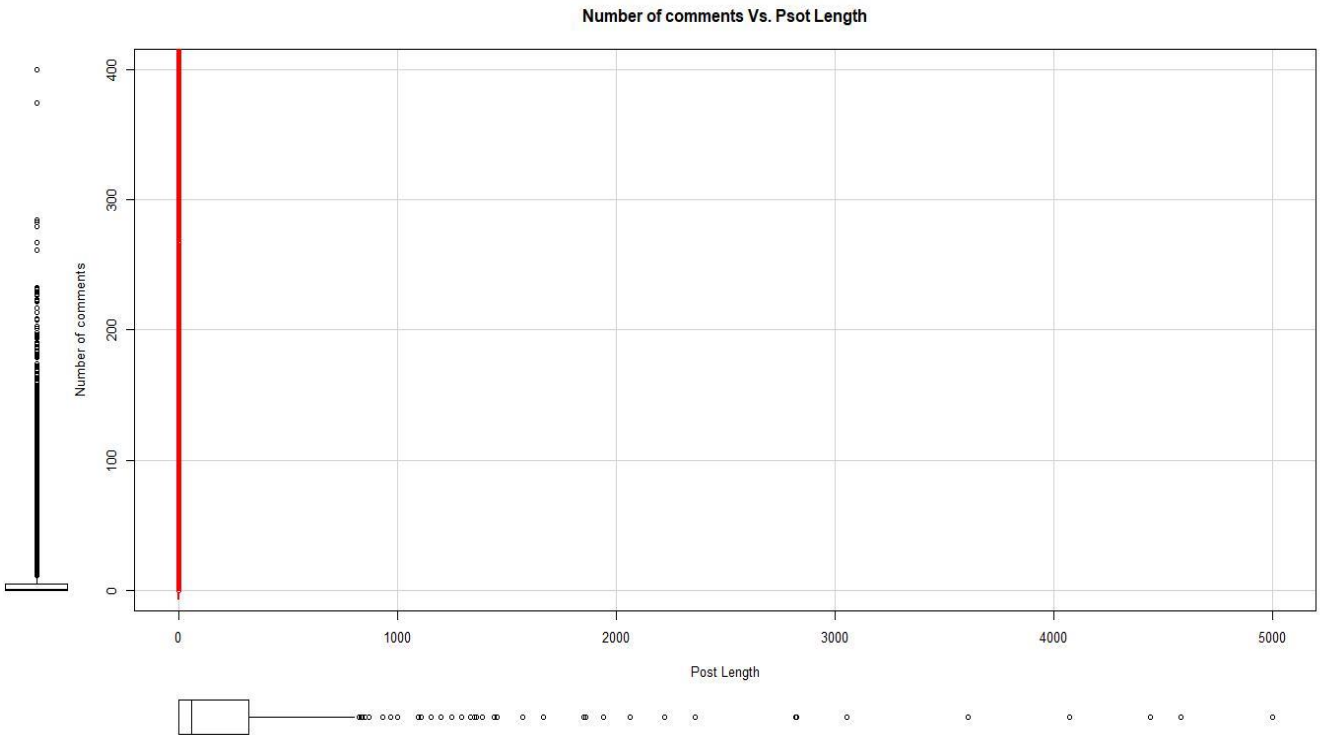
```
> barplot(table(train$target, train$pubday), col = heat.colors(7),
+xlab = "Weekday", ylab = "Number of comments",
+main = "Number of comments Vs. Weekday")
> # post published on Wednesday has maximum comments
```



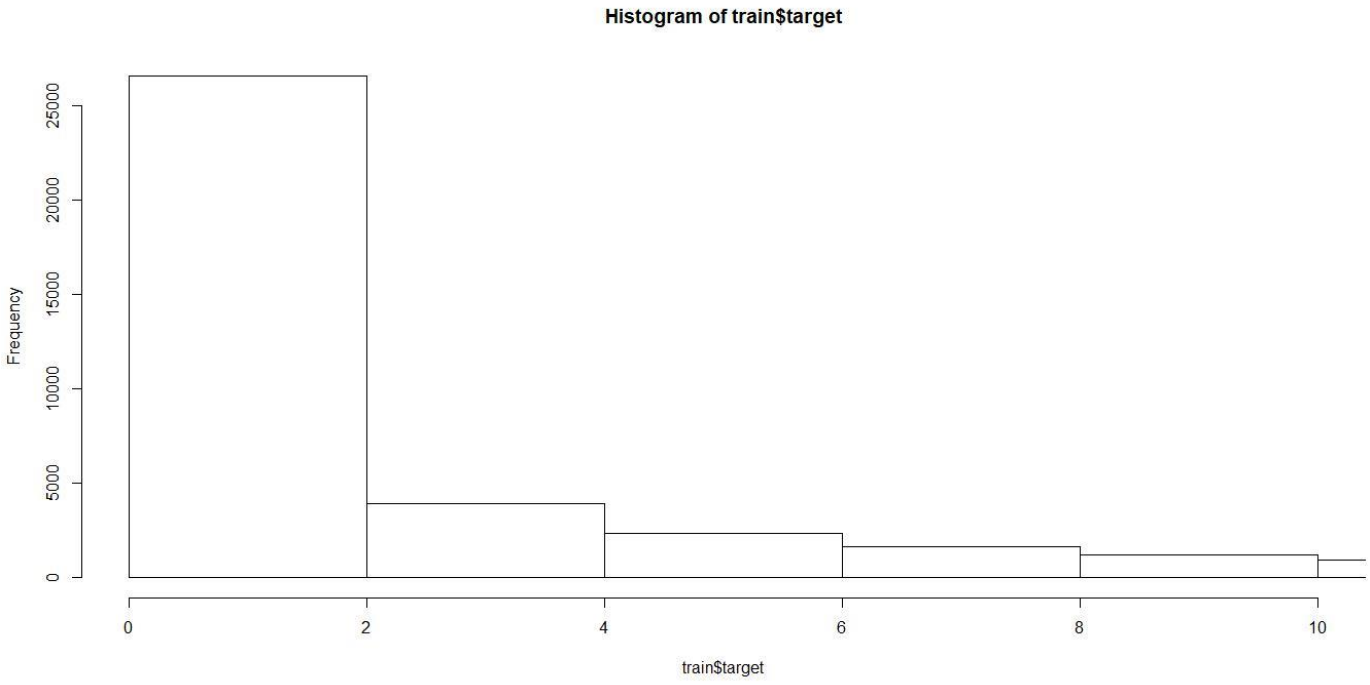
```
> library(car)
Loading required package: carData

Attaching package: ‘car’

The following object is masked from ‘package:dplyr’:

    recode

> # number of comments vs Post Likes
> scatterplot(train$plikes, train$target , col = "Blue",
+             xlab = "Page Likes", ylab = "Number of comments",
+             main =  "Number of commentsVs. Pagelikes",
+             xlim =  c(0,10000000), ylim= c(0,400))
> abline(lm(plikes~target,  data = train), col = "red")
```




```
> # Number of comments Vs Post length
> scatterplot(train$postlength, train$target , col = "Red",
+             xlab = "Post Length", ylab = "Number of comments",
+             main = "Number of comments  Vs. Psot Length",
+             ylim = c(0,400), xlim = c(0,5000))
> abline(lm(postlength~target, data = train), col=  "blue")
```




```
hist(train$target, breaks = 1000, xlim = c(0,10) )
```



Conclusion/Interpretation:

- Posts which are published on Wednesday has maximum comments
- As the page likes increases the comments are not increasing
- As the page length is increasing the number of comments decreases
- Data is very positively skewed. Very less comments after base time

d. Perform any 3 hypothesis tests using columns of your choice, make conclusions.

1. **The R-script for the given problem is as follows:**

```
# Ho: Mean difference bet comments across the publish day is not
significant
 day <- aov(target~pubday, data = train)
summary(day)
```

The output of the R-Script (from Console window) is given as follows:

```
> # Ho: Mean difference bet comments across the publish day is
not significant
> day <- aov(target~pubday, data = train)
> summary(day)
                Df    Sum Sq Mean Sq F value Pr(>F)
pubday           1   7910633 7910633    1221 <2e-16 ***
Residuals    11190 72480187    6477
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
41204 observations deleted due to missingness
```

Conclusion/Interpretation:

Difference between the number of comments after H hrs and comments in first 24 hrs of publish is significant

2. **The R-script for the given problem is as follows:**

```
# Ho: Difference between Mean comments within cc2 and cc4 is not significant
cc2 <- t.test(x=train$cc2, y=train$cc4, paired = FALSE, alternative =
"two.sided", mu=0) cc2
```

The output of the R-Script (from Console window) is given as follows:

```
> # Ho: Difference between Mean comments within cc2 and cc4 is not
  significant
> cc2 <- t.test(x=train$cc2, y=train$cc4, paired = FALSE,
alternative = "two.sided", mu=0)
> cc2

        Welch Two Sample t-test

data:  train$cc2 and train$cc4
t = 122.01, df = 52395, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1885319 0.1946882
sample estimates:
mean of x mean of y
  0.19161   0.00000
```

Conclusion/Interpretation:

Difference between the number of comments in last 24 hrs of base time and comments in first 24 hrs of publish is significant

3.**The R-script for the given problem is as follows:**

```
# Ho: Difference between Mean comments within cc1 and cc3 is not significant
cc3 <- t.test(x=train$cc1, y=train$cc3, paired = FALSE, alternative =
"two.sided", mu=0) cc3
```

The output of the R-Script (from Console window) is given as follows:

```
> cc3 <- t.test(x=train$cc1, y=train$cc3, paired = FALSE,
alternative = "two.sided", mu=0)
> cc3

  Welch Two Sample t-test

data:  train$cc1 and train$cc3
t = -44.255, df = 96439, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2161059 -0.1977756
sample estimates:
mean of x mean of y
0.2791816 0.4861223
```

Conclusion/Interpretation:

Difference between Mean comments within cc1 and cc3 is significant