

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

PART B

(PART B: TO BE COMPLETED BY STUDENTS)

(Students must submit the soft copy as per the following segments within two hours of the practicals. The soft copy must be uploaded on Blackboard LMS or emailed to the concerned Lab in charge Faculties at the end of practical; in case Blackboard is not accessible)

Roll No: C035	Name: Krisha Goti
Class: B	Batch: B1
Date of Experiment: 6/08/2022	Date of Submission
Grade	

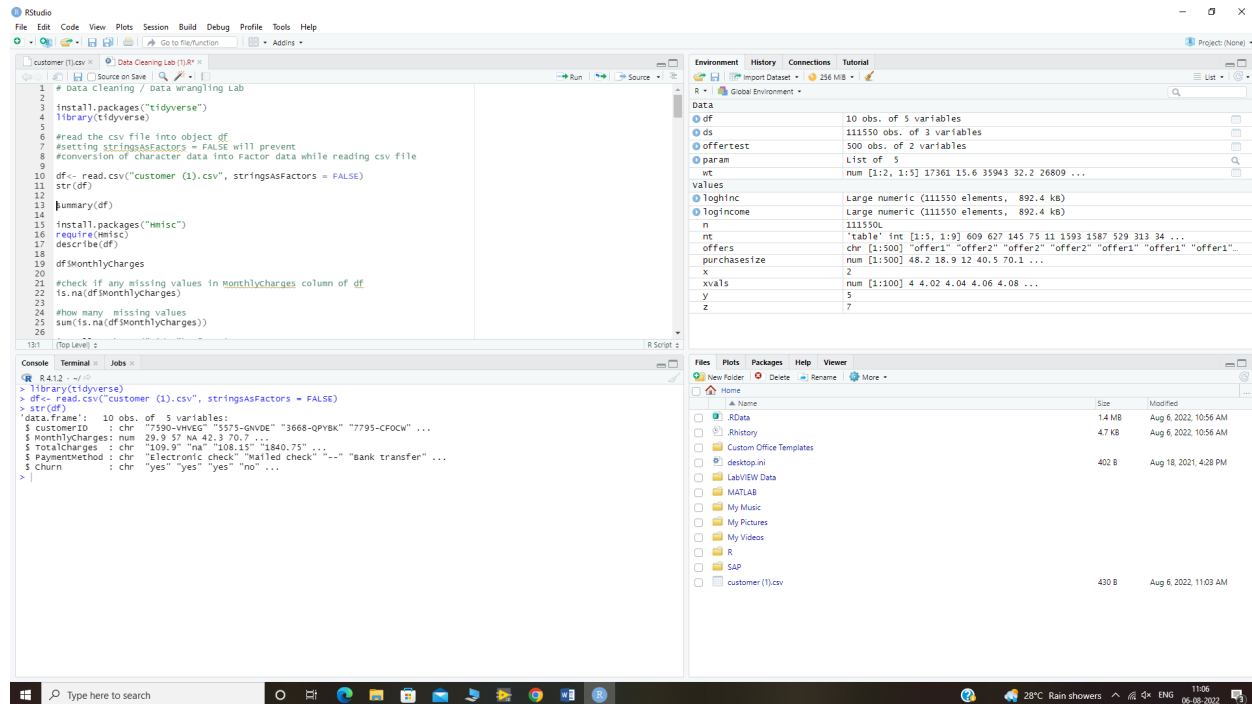
B.1 Work done by student

(Paste your gather information and the comparison table)

1. **Study the working of following commands in R from R documentation by typing them in the 'help' tab**
 - a. `read.csv`
2. **Prepare working environment for the Lab and load data files**
 1. Set the working directory to where we have stored the data.
 2. Read customer.csv dataset using read.csv command
`df<- read.csv("customer.csv", stringsAsFactors = FALSE)`
 3. Display structure of dataframe df
`str(df)`

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



4. Display summary of dataframe to know five point summary of each attribute
5. Install package Harrell Miscellaneous (hmisc)

```
install.packages("Hmisc")
require(Hmisc)
```

This package contains many functions useful for data analysis, high-level graphics, utility operations, functions for computing sample size and power, importing and annotating datasets, imputing missing values, advanced table making, variable clustering, character string manipulation, conversion of R objects to LaTeX and html code, and recoding variables.

`require(package)` load the namespace of the package with name package and attach it on the search list. `require` is designed for use inside other functions; it returns FALSE and gives a warning (rather than an error as `library()` does by default) if the package does not exist.

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

The screenshot shows the RStudio interface. The script editor contains the following R code:

```

1 # data cleaning / data wrangling Lab
2
3 install.packages("tidyverse")
4 library(tidyverse)
5
6 #read the csv file into object df
7 #setting stringsAsFactors = FALSE will prevent
8 #conversion of character data into factor data while reading csv file
9
10 df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
11 str(df)
12
13 summary(df)
14
15 install.packages("Hmisc")
16 require(Hmisc)
17 describe(df)
18
19 #MonthlyCharges
20
21 #check if any missing values in MonthlyCharges column of df
22 is.na(df$MonthlyCharges)
23
24 #how many missing values
25 sum(is.na(df$MonthlyCharges))
26

```

The console output shows the result of the `describe(df)` command:

```

R412 ~> describe(df)
df
 5 variables   10 observations
-----
customerID    n missing distinct
10            0          10

lowest : 1452-KIOVK 3668-QPYBK 5575-GNVDE 6713-OKOMC 7590-VHVEG, highest: 7795-CFOCW 7892-POOKP 8451-AJOMK 9237-HQITU 9305-CD
SKC
value      1452-KIOVK 3668-QPYBK 5575-GNVDE 6713-OKOMC 7590-VHVEG 7795-CFOCW 7892-POOKP 8451-AJOMK 9237-HQITU 9305-CD
SKC
Frequency   1      1      1      1      1      1      1      1      1      1
Proportion  0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1    0.1

MonthlyCharges
  n missing distinct   Info   Mean   Gnd
  7          3          7    1.0  63.97  31.91

lowest : 29.85 42.30 54.10 56.95 70.70, highest: 54.10 56.95 70.70 89.10 104.80
value      29.85 42.30 54.10 56.95 70.70 89.10 104.80
Frequency   1      1      1      1      1      1      1
Proportion  0.143 0.143 0.143 0.143 0.143 0.143 0.143

TotalCharges
  n missing distinct
  9          1          9

```

6. Generate a concise statistical description of dataframe using describe command
7. Display monthlycharges column
8. Check if any missing values are there in column
`is.na(df$MonthlyCharges)`
9. Find how many missing values are there
`sum(is.na(df$MonthlyCharges))`

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

```

5
6 #read the csv file into object df
7 #setting stringsAsFactors = FALSE will prevent
8 #conversion of character data into factor data while reading csv file
9
10 df<- read.csv("customer (1).csv", stringsAsFactors = FALSE)
11 str(df)
12
13 summary(df)
14
15 install.packages("nmisc")
16 require(nmisc)
17 describe(df)
18
19 df$monthlycharges
20
21 #check if any missing values in monthlycharges column of df
22 is.na(df$monthlycharges)
23
24 #how many missing values
25 sum(is.na(df$monthlycharges))
26
27 install.packages("tidyr") #for pipe operator
28 library(tidyr)
29
30 install.packages("dplyr") #for distinct function
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

10. Install package tidyr and dplyr

```

install.packages("tidyr") #for pipe operator
library(tidyr)
install.packages("dplyr") #for distinct function
library(dplyr)

```

11. find unique values in MonthlyCharges column

```
df %>% distinct(MonthlyCharges)
```

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains R code for data cleaning and summarization. The code includes comments and functions for handling missing values, installing packages (tidyr, dplyr), and summarizing data by distinct values.
- Environment:** Displays the current data environment, showing variables like 'df', 'ds', 'offerfest', 'param', and 'wt' with their respective data types and sizes.
- Console:** Shows the output of the R code, including the results of the 'df %>% distinct(MonthlyCharges)' command, which lists the distinct values of the 'MonthlyCharges' column.
- Files:** Shows the file structure of the project, including the 'Data Cleaning Lab (1).R' file.

12. summarise distinct values

```
df %>% summarise(n= n_distinct(MonthlyCharges))
```

‘n_distinct’ Efficiently count the number of unique values in a set of vector. **This is a faster and more concise equivalent of length(unique(x))**

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains R code for data cleaning and summarization. The code includes comments and functions to handle missing values, install packages (tidyr, dplyr), and summarize data by distinct values in the 'MonthlyCharges' column.
- Environment:** Lists the loaded data frames: 'df' (10 obs. of 5 variables), 'ds' (111550 obs. of 3 variables), 'offerTest' (500 obs. of 2 variables), 'param' (List of 5), and 'wt' (num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...).
- Console:** Shows the output of the R code, including the results of the 'distinct' and 'summarise' functions.
- Files:** Lists the installed packages and their versions, including 'askpass', 'assertthat', 'backports', 'base', 'base64enc', 'bit', 'bit64', 'blob', 'boot', 'brglm2', 'broom', 'cachem', 'callr', 'cellranger', 'checkmate', 'class', 'cli', 'clipr', and 'cluster'.

13. doing multiple things in summarise

```
df%>% summarise(n=n_distinct(MonthlyCharges),
                 count = sum(is.na(MonthlyCharges)),
                 M = mean(MonthlyCharges, na.rm=TRUE))
```

14. replace missing values with median

```
df <- df %>% mutate(MonthlyCharges
                    =replace(MonthlyCharges,is.na(MonthlyCharges),median(MonthlyCharges,na.rm = TRUE)))
```

15. checking for nonstandard missing values:

```
is.na(df$TotalCharges) #detects only single null value
df%>% summarise(n=sum(is.na(TotalCharges)))
```

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains R code for cleaning the 'MonthlyCharges' column. The code includes:


```

48 df$%>% summarise(n=n_distinct(MonthlyCharges),
49 count = sum(is.na(MonthlyCharges))),
50 m = mean(MonthlyCharges, na.rm=TRUE))
51
52 #replace missing values with median
53
54 df <- df %>% mutate(MonthlyCharges
55 ~replace(MonthlyCharges, is.na(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))
56
57 #checking for nonstandard missing values
58
59 is.na(df$totalCharges) #detects only single null value
60
61 df$%>% summarise(n=sum(is.na(totalCharges)))
62
63 #change the 'na' and 'N/A' values to NA in totalCharges column
64 df <- df %>%
65 mutate(totalCharges = replace(totalCharges, totalCharges == "na", NA)) %>%
66 mutate(totalCharges = replace(totalCharges, totalCharges == "N/A", NA))
67
68 is.na(df$totalCharges) # now the count is 3
69 df$totalCharges
70
71 str(df)
72 #convert totalCharges to Numeric
73 df$totalCharges <- as.numeric(df$totalCharges)
      
```
- Environment:** Lists variables in the global environment:
 - df: 10 obs. of 5 variables
 - ds: 111550 obs. of 3 variables
 - offerTest: 500 obs. of 2 variables
 - param: List of 5
 - wt: num [1:2, 1:5] 17361 15.6 35943 32.2 26809 ...
 - values:
 - logIncome: Large numeric (111550 elements, 892.4 kb)
 - n: Large numeric (111550 elements, 892.4 kb)
 - rt: 111550L
 - table: int [1:5, 1:9] 609 627 145 75 11 1593 1587 529 313 34 ...
 - offers: chr [1:500] "offer1" "offer2" "offer2" "offer2" "offer2" "offer1" "offer1" "offer1"
 - purchasesize: num [1:500] 48.2 18.9 12 40.5 70.1 ...
 - x: 2
 - xvals: num [1:100] 4 4.02 4.04 4.06 4.08 ...
 - y: 5
 - z: 7
- Console:** Shows the output of the R code:

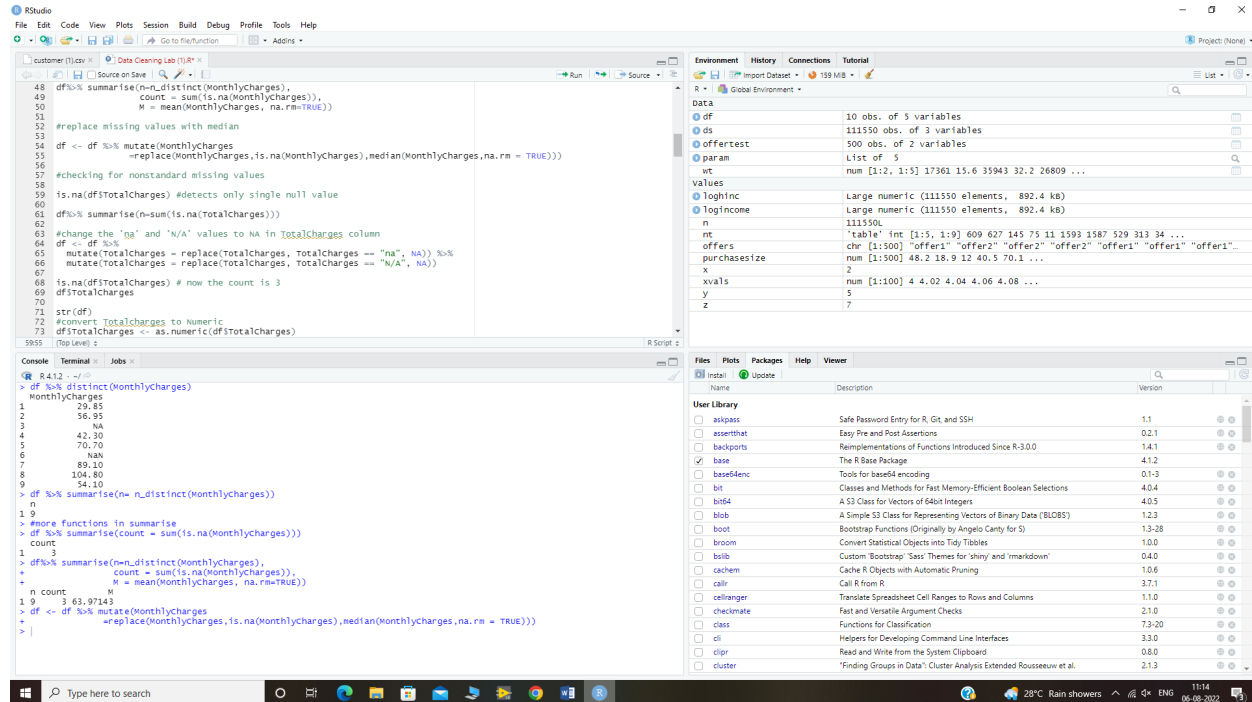

```

R 4.1.2 ~ /
> df %>% distinct(MonthlyCharges)
MonthlyCharges
1      29.85
2      56.95
3      NA
4      42.30
5      70.70
6      NA
7      89.10
8      104.80
9      54.10
> df %>% summarise(n= n_distinct(MonthlyCharges))
n
1 9
> #more functions in summarise
> df %>% summarise(count = sum(is.na(MonthlyCharges)))
count
1 3
> df$%>% summarise(m=m_distinct(MonthlyCharges),
+ count = sum(is.na(MonthlyCharges)),
+ m = mean(MonthlyCharges, na.rm=TRUE))
n count
1 9 3 63.97143
> df <- df %>% mutate(MonthlyCharges
+ ~replace(MonthlyCharges, is.na(MonthlyCharges), median(MonthlyCharges, na.rm = TRUE)))
+
      
```
- Files:** Shows the installed packages and user library.

16. change the 'na' and 'N/A' values to NA in TotalCharges column. Then count and display null values in totalCharges column
17. Display all values in totalCharges column

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23



18. convert Totalcharges to Numeric using as.numeric command

19. Describe structure of dataframe

20. replace the missing values with mean value in Totalcharges column and display totalcharges column. Ignore null values while calculating mean value.

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

The screenshot shows the RStudio IDE with the following components:

- Source Editor:** Contains R code for cleaning the 'TotalCharges' column. It replaces 'na' and 'N/A' with NA, converts the column to numeric, and replaces missing values with the mean of the non-missing values.
- Environment Pane:** Shows the current data environment with variables: df (10 obs. of 5 variables), ds (111550 obs. of 3 variables), offerTest (500 obs. of 2 variables), param (List of 5), vt (num [1:2, 1:5]), logInc (Large numeric), n (111550L), mt (matrix), offers (chr), purchaseSize (num), xvals (num), y (5), and z (7).
- Console:** Shows the execution of the R code, including the conversion of 'TotalCharges' to numeric and the replacement of missing values.
- Files Pane:** Shows the project files and folders.

21. check 'paymentmethod' column for null values and comment on the result.
22. Replace '-' by 'NA' and null value by 'unavailable'
23. Add new column 'percentagecharges' using 'mutate' command

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

```

95 is.na(df$PaymentMethod)
96
97 df <- df %>%
98   mutate(PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "unavailable"))
99
100 #using mutate to add new column
101
102 df <- df %>%
103   mutate(PercentageCharges= MonthlyCharges*100/ TotalCharges)
104
105 #Subsetting data
106 #Subsetting rows using filter
107 #dplyr:: filter(iris, Sepal.Length > 7)
108
109 data[["iris"]]
110
111 str(iris)
112 levels(iris$species)
113
114 #Extract rows that meet logical criteria Sepal.Length > 7.
115
116 large_sepal <- filter(iris, iris$Sepal.Length>7)
117
118 large_sepal
119
120 #or
121
122 iris[iris$Sepal.Length>7,]

```

```

R 4.1.2 ~ /
> data.frame() 10 obs. of 5 variables:
> $ customerID: chr "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
> $ MonthlyCharges: num 29.9 57.57 42.3 70.7 ...
> $ TotalCharges: num 110 NA 108 1841 NA ...
> $ PaymentMethod: chr "Electronic check" "Mailed check" "--" "Bank transfer" ...
> $ Churn: chr "yes" "yes" "yes" "no" ...
> #replace the missing values with mean value in TotalCharges column
> df <- df %>%
+   mutate(TotalCharges
+     = replace(TotalCharges,
+               is.na(TotalCharges),
+               mean(TotalCharges, na.rm = TRUE)))
> df$TotalCharges
[1] 109.900 1175.671 108.150 1840.750 1175.671 820.500 1949.400 1175.671 3046.050 354.950
> #check PaymentMethod column
> df$PaymentMethod
[1] "Electronic check" "Mailed check" "--" "Bank transfer" "Electronic check" "--"
[7] "Credit card"
> is.na(df$PaymentMethod) #all values are false here which should not be the case
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> df <- df %>%
+   mutate(PaymentMethod = replace(PaymentMethod, PaymentMethod == "--", NA))
> is.na(df$PaymentMethod)
[1] FALSE FALSE TRUE FALSE FALSE TRUE FALSE FALSE FALSE
> df <- df %>%
+   mutate(PaymentMethod = replace(PaymentMethod, is.na(PaymentMethod), "unavailable"))
> df <- df %>%
+   mutate(PercentageCharges= MonthlyCharges*100/ TotalCharges)

```

24. Display dataframe

	customerID	MonthlyCharges	TotalCharges	PaymentMethod	Churn	PercentageCharges
1	7590-VHVEG	29.85	109.900	Electronic check	yes	27.161056
2	5575-GNVDE	56.95	1175.671	Mailed check	yes	4.844040
3	3668-QPYBK	56.95	108.150	unavailable	yes	52.658345
4	7795-CFOCW	42.30	1840.750	Bank transfer	no	2.297976
5	9237-HQITU	70.70	1175.671	Electronic check	no	6.013585
6	9305-CDSKC	56.95	820.500	unavailable	yes	6.940890
7	1452-KIOVK	89.10	1949.400	Credit card	no	4.570637
8	6713-OKOMC	56.95	1175.671		yes	4.844040
9	7892-POOKP	104.80	3046.050	Electronic check	no	3.440521
10	8451-AJOMK	54.10	354.950	Electronic check	no	15.241583

B.2 Conclusion

SVKM'S NMIMS Deemed-to-be-University
Mukesh Patel School of Technology Management & Engineering
Department of Computer Engineering

Program	BTech Intg.	Branch	Computers
Semester	IX	Year	V
Name of the Faculty	Artika Singh	Class	Div B and Div C
Course Title	Data Mining	Academic year	2022-23

After completing this experiment, I am able to Apply appropriate data cleaning techniques and improve data quality and to make it complete and consistent.