



American International University-Bangladesh (AIUB)

**Department of Computer Science
Faculty of Science & Technology (FST)**

INTRODUCTION TO DATA SCIENCE

Report

Submitted By

Semester: Fall_24_25	
Student Name	Student ID
Md.Shariar Hosain Sanny	21-44677-1
Redoanul Haque Riyad	21-44708-1
Shariar Rahman Apurbo	21-44738-1
Nafis Mubasshir Shah	21-45158-2

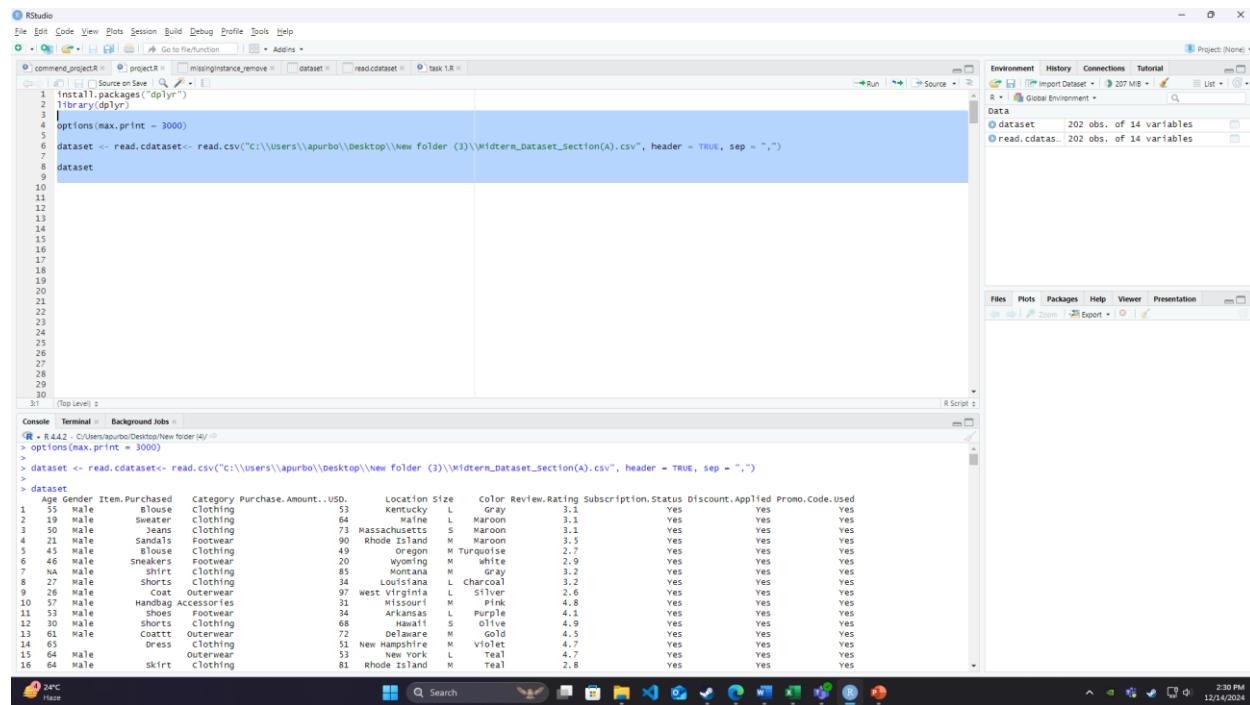
Serial Number	Task	Page Number
1	Short Description	3
1	Loading and Displaying Dataset	3
2	Find and handle the missing values:	3-8
3	Missing value in graph	9
4	Find and remove duplicate values.	9-10
5	Convert the imbalanced data set into the balanced data set	11
6	filtering methods	12-13
7	Convert attributes from numeric to categorical	13-14
8	Convert attributes from categorical to numeric	14-15
9	Outliers	15-16
10	Normalization method	18

Short Description

The dataset contains detailed transactional information about individual purchases, including attributes such as Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Review Rating, Subscription Status, Discount Applied, and Promo Code Used. It also includes behavioral metrics like Previous Purchases and Frequency of Purchases. The data features a mix of numerical values and boolean indicators, making it versatile for analysis. This dataset is labeled, as it includes target variables suitable for predictive analysis, and is designed for use in supervised learning tasks.

Loading and Displaying Dataset :

First we import the dataset using `read.csv` then we view the dataset by calling its variable.



Find and handle the missing values:

> dataset	Age	Gender	Item.Purchased	Category	Purchase.Amount., USD.	Location	Size	Color	Review.Rating	Subscription.Status	Discount	Applied	Promo.Code	Used
1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	3.1	Yes	Yes	Yes	Yes	Yes
2	19	Male	Sweater	Clothing	64	Maine	M	Maroon	3.1	Yes	Yes	Yes	Yes	Yes
3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	3.1	Yes	Yes	Yes	Yes	Yes
4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	3.5	Yes	Yes	Yes	Yes	Yes
5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	2.7	Yes	Yes	Yes	Yes	Yes
6	46	Male	Sneakers	Footwear	20	Wyoming	M	White	2.9	Yes	Yes	Yes	Yes	Yes
7	NA	Male	Shirt	Clothing	85	Montana	M	Gray	3.2	Yes	Yes	Yes	Yes	Yes
8	27	Male	Shorts	Clothing	34	Louisiana	L	Charcoal	3.2	Yes	Yes	Yes	Yes	Yes
9	26	Male	Coat	Outerwear	97	West Virginia	L	Silver	2.6	Yes	Yes	Yes	Yes	Yes
10	57	Male	Handbag	Accessories	31	Missouri	M	Pink	4.8	Yes	Yes	Yes	Yes	Yes
11	53	Male	Shoes	Footwear	34	Arkansas	L	Purple	4.1	Yes	Yes	Yes	Yes	Yes
12	30	Male	Shorts	Clothing	68	Hawaii	S	Olive	4.9	Yes	Yes	Yes	Yes	Yes
13	61	Male	Coat	Outerwear	72	Delaware	M	Gold	4.5	Yes	Yes	Yes	Yes	Yes
14	65	Male	Dress	Clothing	51	New Hampshire	M	Violet	4.7	Yes	Yes	Yes	Yes	Yes
15	64	Male	Outerwear	Clothing	53	New York	L	Teal	4.7	Yes	Yes	Yes	Yes	Yes
16	64	Male	Skirt	Clothing	81	Rhode Island	M	Teal	2.8	Yes	Yes	Yes	Yes	Yes

Some of the data are empty so we must first convert them to NA. We convert empty data using `dataset[dataset == ""] <- NA`

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and Addins. The top-left pane displays a script named 'command.R' with code for dataset manipulation. The top-right pane shows the Environment browser with 'dataset' and 'read.cdataset' objects. The bottom-left pane is the Console, showing the command '> dataset[dataset == ""] <- NA' being run. The bottom-right pane is the Global History browser. The bottom of the screen features the Windows taskbar with various pinned icons.

Then we use `is.na(dataset)` to find the missing values in the dataset `sum(is.na(dataset))` to get the total count of missing values and `colSums(is.na(dataset))` to count the missing values in each column.

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project:** Project "R" is selected.
- Code Editor:** The script pane contains R code for handling missing values. Lines 32-33 show the use of `sum(is.na(dataset))` and `colSums(is.na(dataset))` to identify missing values. Line 36 shows the assignment of `is.na(dataset)` to a variable. Lines 37-43 show the use of `sum(is.na(dataset))` and `colSums(is.na(dataset))` again.
- Console:** The console output shows the execution of the R code. It starts with `R 4.4.2 - C:/Users/tpurbo/Desktop/keras/folder [4] <` followed by the results of the missing value checks. Then, it shows the creation of a dataset with columns: Age, Gender, Item.Purchased, Category, Purchase.Amount..USD, Location, Size, Color, Review.Rating, Subscription.Status, Discount.Applied, Promo.Code.Used, Previous.Purchases, and Frequency.of.Purchases. The final command `> is.na(dataset)` is shown, followed by the first six rows of the dataset.
- Data View:** The data pane shows the `dataset` and `read.cdatass` objects. Both have 202 observations and 14 variables.
- Plots:** A histogram of a variable is visible in the plots pane.
- Environment:** The environment pane shows the global environment with variables like `dataset` and `read.cdatass`.
- Help:** The help pane is visible at the bottom.

We find the positions of missing values

```

41
42
43
44
45
46
47 which(is.na(dataset$age))
48 which(is.na(dataset$Gender))
49 which(is.na(dataset$item.Purchased))
50 which(is.na(dataset$Frequency.of.Purchases))
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71 (Top Level) z

```

```

Console Terminal Background Jobs
(R - R 4.4.2 - C:\Users\zhuo\Desktop\New folder (4)\)
> which(is.na(dataset$age))
[1] 7 25 45 139
> which(is.na(dataset$Gender))
[1] 14 48 165 170
> which(is.na(dataset$item.Purchased))
[1] 15 44
> which(is.na(dataset$Frequency.of.Purchases))
[1] 9 18
>

```

We use two methods handle the missing values

1. Discard Instances
2. Replace by Most Frequent/Average Value

Discard Instances:

Here we handle missing value by remove it

```

36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52 missingInstance_remove <- na.omit(dataset)
53 missingInstance_remove
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71 (Top Level) z

```

```

Console Terminal Background Jobs
(R - R 4.4.2 - C:\Users\zhuo\Desktop\New folder (4)\)
> missingInstance_remove <- na.omit(dataset)
>
> missingInstance_remove
Age Gender Item.Purchased Category Purchase.Amt..USD. Location Size Color Review.Rating Subscription.Status Discount.Applied Promo.Code.Used
1 55 Male Blouse Clothing 53 Kentucky L Gray 3.1 Yes Yes Yes
2 19 Male Sweater Clothing 54 Maine L Maroon 3.1 Yes Yes Yes
3 34 Male Shorts Clothing 73 Massachusetts S Maroon 3.1 Yes Yes Yes
4 21 Male Sandals Footwear 90 Rhode Island M Maroon 3.5 Yes Yes Yes
5 45 Male Blouse Clothing 49 Oregon M Turquoise 2.7 Yes Yes Yes
6 47 Male Sneakers Footwear 20 Wisconsin White 2.0 Yes Yes Yes
8 27 Male Shorts Clothing 34 Louisiana L Charcoal 3.2 Yes Yes Yes
10 57 Male Handbag Accessories 31 Missouri M Pink 4.8 Yes Yes Yes
11 53 Male Shoes Footwear 34 Arkansas L Purple 4.1 Yes Yes Yes
12 39 Male Shorts Clothing 68 Connecticut S Blue 4.9 Yes Yes Yes
13 61 Male Coats Clothing 72 Delaware M Gold 4.5 Yes Yes Yes
16 64 Male Skirt Clothing 81 Rhode Island M Teal 2.8 Yes Yes Yes
17 52 Male Sunglasses Accessories 38 New Jersey S Gray 4.1 Yes Yes Yes
19 52 Male Sweater Clothing 48 Montana S Black 4.6 Yes Yes Yes
20 66 Male Pants Clothing 90 Rhode Island M Green 3.3 Yes Yes Yes
21 21 Male Pants Clothing 51 Louisiana M Black 2.8 Yes Yes Yes
22 56 Male Pants Clothing 62 North Carolina M Charcoal 4.1 Yes Yes Yes
23 56 Male Pants Clothing 37 California M Peach 3.2 Yes Yes Yes

```

No missing values found.

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
File -> project -> missinginstance_remove <- dataset <- read.dataset <- task 1.R
Source | Run | Source | History | Environment | Data | Files | Plots | Packages | Help | Viewer | Presentation
38
39
40
41
42
43
44
45
46
47
48
49 sum(is.na(missinginstance_remove))
50 colSums(is.na(missinginstance_remove))
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
49: [Top Level] z
Console Terminal Background Jobs
R 4.4.2 - C:\Users\spurbo\Desktop\New folder (4)\missinginstance_remove.R
190      4   Fortnightly
191      2   annually
192     28   bi-weekly
193     24   Annually
194     14   Fortnightly
195     11   weekly
196     27   bi-weekly
197      7   Fortnightly
198     49   annually
199     21   annually
200     10   Every 3 Months
201     37   weekly
202     9   Every 3 Months
> sum(is.na(missinginstance_remove))
[1] 0
> colSums(is.na(missinginstance_remove))
     Age    gender  Item.Purchased Category Purchase.Amount..USD. Location    size
       0        0           0          0            0          0        0
Color Review.Rating Subscription.Status Discount.Applied Promo.Code.Used Previous.Purchases Frequency.of.Purchases
       0        0           0          0            0          0        0
> |
24°C Haze
3:17 PM 12/14/2024

```

Replace by Most Frequent/Average Value

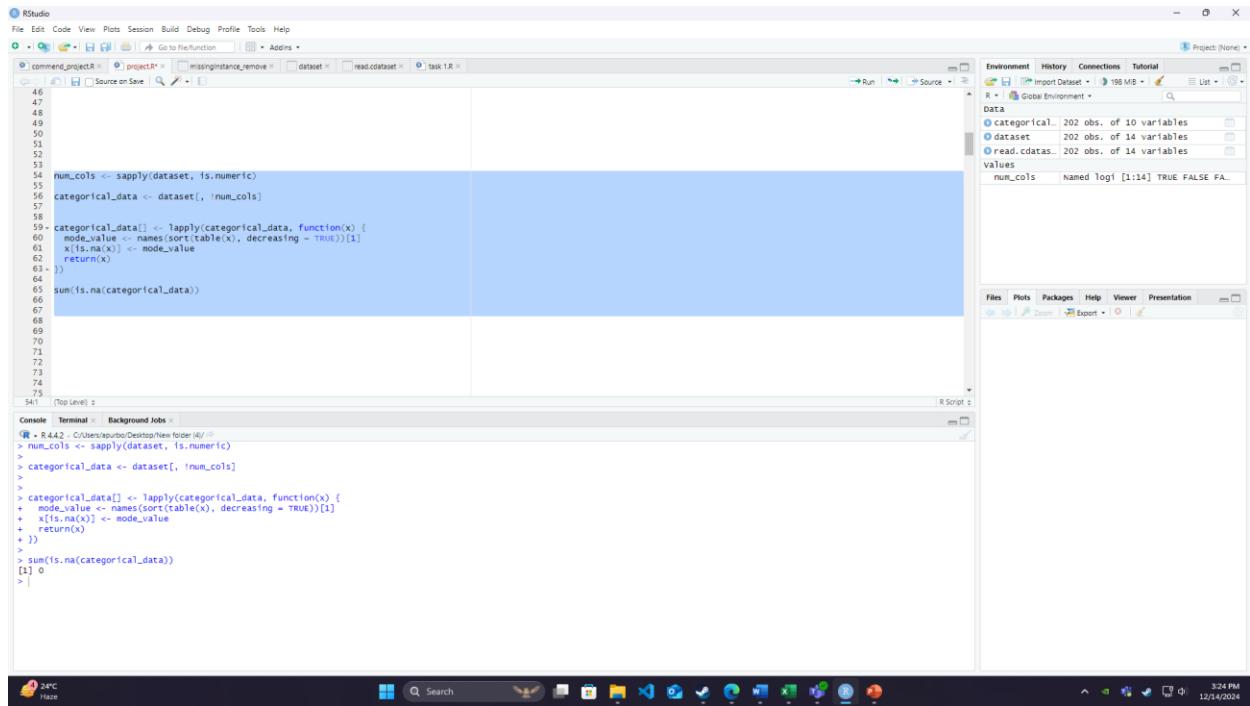
First, we detect categorical and numerical attributes. If the result is true the attribute is numerical, if False it is categorical

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
File -> project -> missinginstance_remove <- dataset <- read.dataset <- task 1.R
Source | Run | Source | History | Environment | Data | Files | Plots | Packages | Help | Viewer | Presentation
37
38
39
39
40
41
42
43
44 num_cols <- sapply(dataset, is.numeric)
45 num_cols
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
43: [Top Level] z
Console Terminal Background Jobs
R 4.4.2 - C:\Users\spurbo\Desktop\New folder (4)\missinginstance_remove.R
> num_cols <- sapply(dataset, is.numeric)
> num_cols
     Age    gender  Item.Purchased Category Purchase.Amount..USD. Location    size
       0        0           0          0            0          0        0
Color Review.Rating Subscription.Status Discount.Applied Promo.Code.Used Previous.Purchases Frequency.of.Purchases
       0        0           0          0            0          0        0
> |
24°C Haze
3:59 PM 12/14/2024

```

In the case of categorical attributes, we handle missing values by replacing them with the most frequent value (mode) in each instance.



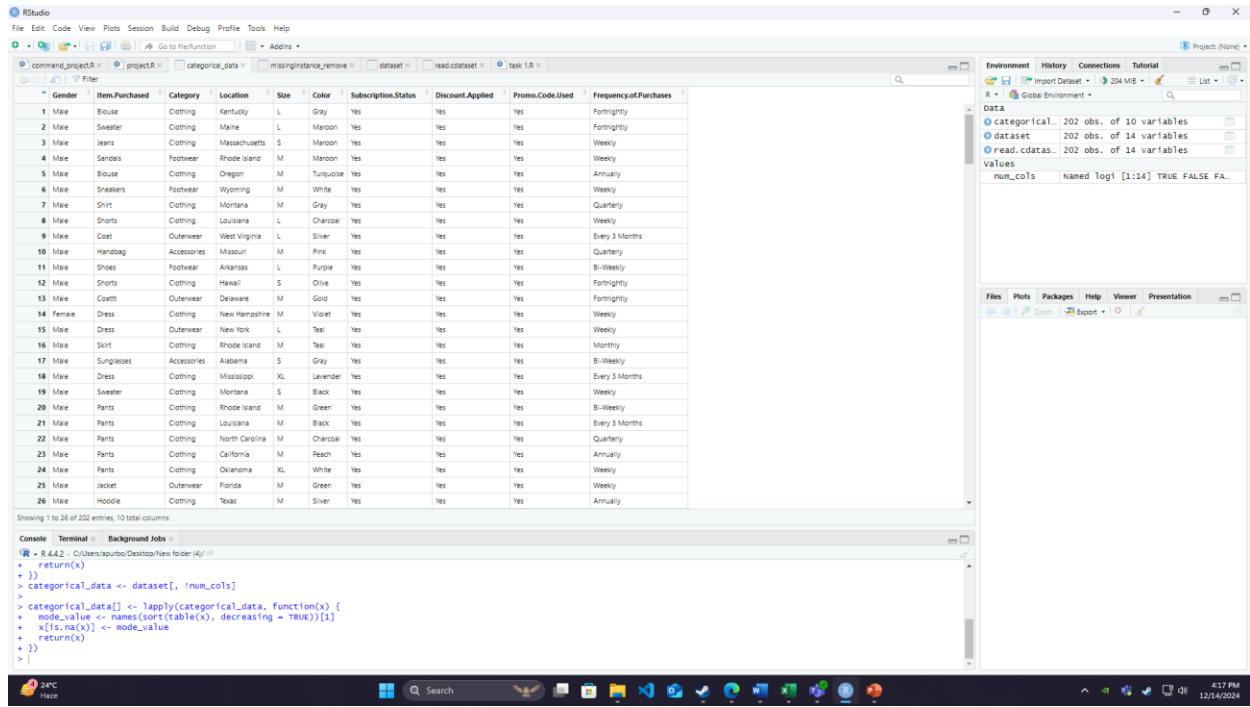
```

46
47
48
49
50
51
52
53
54 num_cols <- sapply(dataset, is.numeric)
55 categorical_data <- dataset[, !num_cols]
56
57
58 categorical_data[] <- lapply(categorical_data, function(x) {
+   mode_value <- names(sort(table(x), decreasing = TRUE))[1]
+   x[is.na(x)] <- mode_value
+   return(x)
+ })
59
60 sum(is.na(categorical_data))
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75

```

The screenshot shows the RStudio interface with the code editor containing the provided R script. The code uses `sapply` to find numeric columns and then `lapply` to iterate over the remaining categorical columns. For each column, it calculates the mode using `names(sort(table(x), decreasing = TRUE))[1]` and replaces missing values (NA) with this mode. The script then sums the number of missing values across all categorical columns.

Replacing them with the most frequent value (mode) in each instance.



	Gender	Item.Purchased	Category	Location	Size	Color	Subscription.Status	Discount.Applied	Promo.Code.Used	Frequency.of.Purchases
1	Male	Bouse	Clothing	Kentucky	L	Gray	Yes	Yes	Yes	Formerly
2	Male	Sweater	Clothing	Maine	L	Maroon	Yes	Yes	Yes	Formerly
3	Male	Jean	Clothing	Massachusetts	S	Maroon	Yes	Yes	Yes	Weekly
4	Male	Sandals	Footwear	Rhode island	M	Maroon	Yes	Yes	Yes	Weekly
5	Male	Bouse	Clothing	Oregon	M	Turquoise	Yes	Yes	Yes	Annually
6	Male	Sneakers	Footwear	Wyoming	M	White	Yes	Yes	Yes	Weekly
7	Male	Shirt	Clothing	Montana	M	Gray	Yes	Yes	Yes	Quarterly
8	Male	Shorts	Clothing	Louisiana	L	Charcoal	Yes	Yes	Yes	Weekly
9	Male	Coat	Outerwear	West Virginia	L	Silver	Yes	Yes	Yes	Every 3 Months
10	Male	Handbag	Accessories	Missouri	M	Pink	Yes	Yes	Yes	Quarterly
11	Male	Shoes	Footwear	Arkansas	L	Purple	Yes	Yes	Yes	Bi-Weekly
12	Male	Shorts	Clothing	Hawaii	S	Onyx	Yes	Yes	Yes	Formerly
13	Male	Coat	Outerwear	Delaware	M	Gold	Yes	Yes	Yes	Formerly
14	Female	Dress	Clothing	New Hampshire	M	Violet	Yes	Yes	Yes	Weekly
15	Male	Dress	Outerwear	New York	L	Teal	Yes	Yes	Yes	Weekly
16	Male	Skirt	Clothing	Rhode island	M	Teal	Yes	Yes	Yes	Monthly
17	Male	Sunglasses	Accessories	Alabama	S	Gray	Yes	Yes	Yes	Bi-Weekly
18	Male	Dress	Clothing	Mississippi	XL	Lavender	Yes	Yes	Yes	Every 3 Months
19	Male	Sweater	Clothing	Montana	S	Black	Yes	Yes	Yes	Weekly
20	Male	Pants	Clothing	Rhode island	M	Green	Yes	Yes	Yes	Bi-Weekly
21	Male	Pants	Clothing	Louisiana	M	Black	Yes	Yes	Yes	Every 3 Months
22	Male	Pants	Clothing	North Carolina	M	Charcoal	Yes	Yes	Yes	Quarterly
23	Male	Pants	Clothing	California	M	Peach	Yes	Yes	Yes	Annually
24	Male	Pants	Clothing	Oklahoma	XL	White	Yes	Yes	Yes	Weekly
25	Male	Jacket	Outerwear	Florida	M	Green	Yes	Yes	Yes	Weekly
26	Male	Hoodie	Clothing	Texas	M	Silver	Yes	Yes	Yes	Annually

The screenshot shows the RStudio interface with the modified dataset. The `dataset` variable now contains 202 observations and 14 variables. The data frame includes columns for gender, item purchased, category, location, size, color, subscription status, discount applied, promo code used, and frequency of purchases. The console shows the execution of the R script, which replaces missing values in the categorical data with their respective modes.

In the case of numerical attributes, we handle missing values by replacing them with the most average value (mean) in each instance.

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Displays R code for handling missing values in a dataset. The code uses `lapply` to iterate over numerical columns, calculate the mean, and replace missing values (NA) with the mean. It also rounds the age column.
- Terminal:** Shows the command-line history of the R session, including the execution of the script and its output.
- Environment:** Shows the global environment with objects like `dataset`, `numerical_data`, and `values`.
- Plots:** No plots are visible in this screenshot.
- Session:** Shows the current R version (R 4.4.2) and the working directory (C:/Users/abubu/Desktop/New folder/4/).
- System:** Shows the Windows taskbar with various pinned icons.

Replacing them with the most average value (mean) in each instance.

The screenshot shows the RStudio interface with the following details:

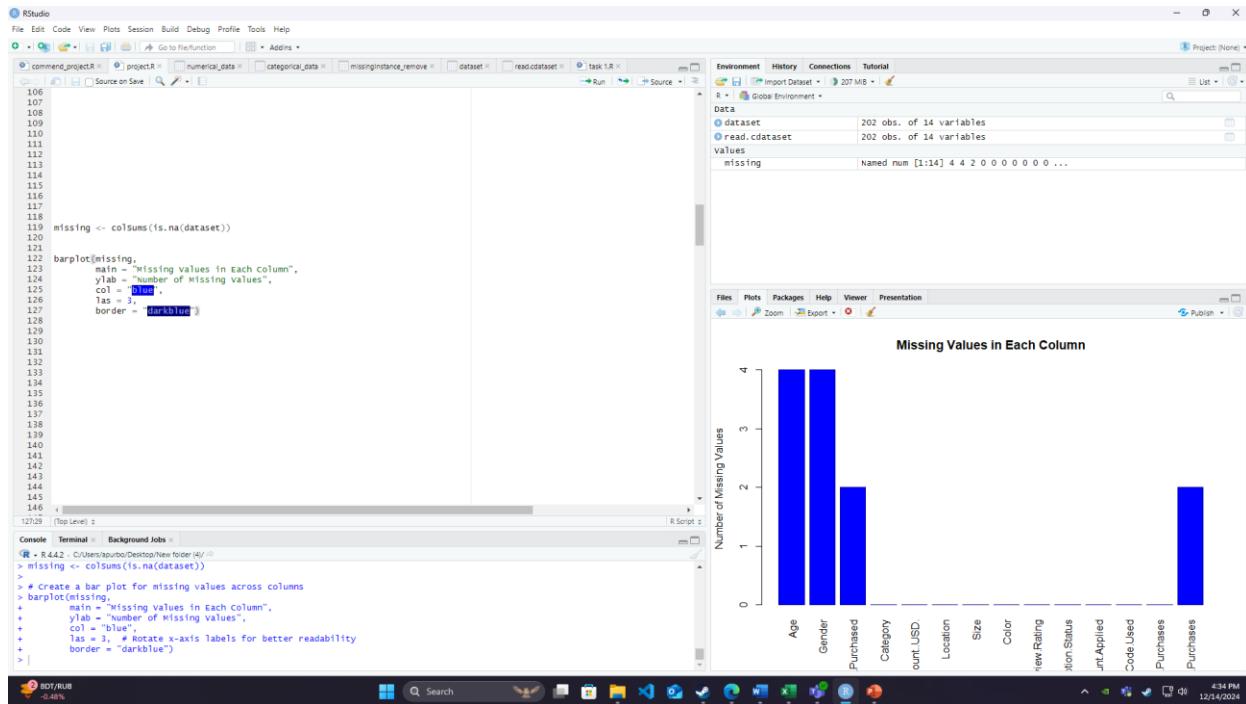
- Data View:** Displays the modified dataset `numerical_data` as a table. The first few rows are:

	Age	Purchase.Amount..USD	Review.Rating	Previous.Purchases
1	55	53	3.1	14
2	19	64	3.1	2
3	50	73	3.1	23
4	21	90	3.5	49
5	45	49	2.7	31
6	46	20	2.9	14
7	48	85	3.2	49
8	27	34	3.2	19
9	26	97	2.6	8
10	57	31	4.0	4
11	53	34	4.1	26
12	30	68	4.9	10
13	61	72	4.5	37
14	65	51	4.7	31
15	64	53	4.7	34
16	64	81	2.8	8
17	25	36	4.1	44
18	53	38	4.7	36
19	52	48	4.6	17
20	66	90	3.3	46
21	21	51	2.6	50
22	31	62	4.1	22
23	56	37	3.2	32
24	31	88	4.4	40
25	46	22	2.9	16
26	18	25	3.6	14

- Terminal:** Shows the command-line history of the R session, including the execution of the script and its output.
- Environment:** Shows the global environment with objects like `dataset`, `numerical_data`, and `values`.
- Plots:** No plots are visible in this screenshot.
- Session:** Shows the current R version (R 4.4.2) and the working directory (C:/Users/abubu/Desktop/New folder/4/).
- System:** Shows the Windows taskbar with various pinned icons.

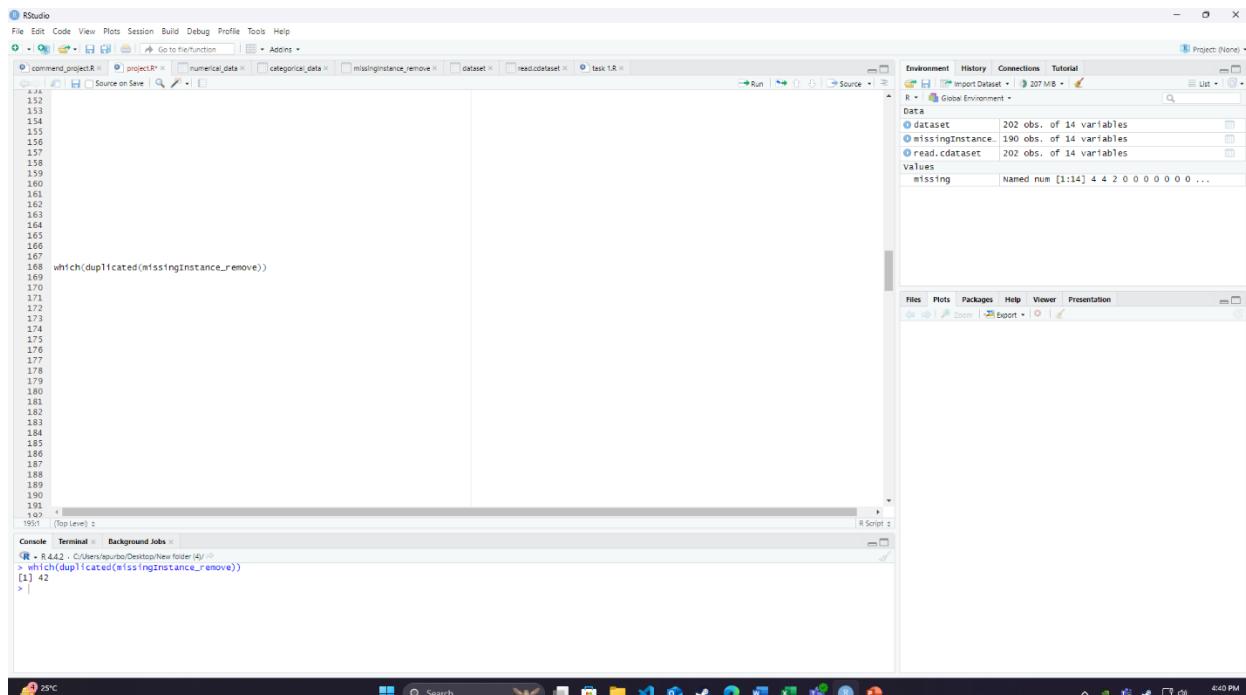
Missing value in graph

We use barplot for the graph



Find and remove duplicate values.

Find and identify the positions of duplicate value.



Remove duplicate and show the output

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Displays the command `duplicate_value <- distinct(missingInstance_remove)`.
- Console:** Shows the command `> which(duplicated(duplicate_value))` being run.
- Environment View:** Shows the variable `dataset` with 202 observations and 14 variables.
- Output View:** Shows the results of the `which` command, indicating no duplicates found.
- System Tray:** Shows the date and time as 12/14/2024 at 4:48 PM.

The screenshot shows the RStudio interface with the following details:

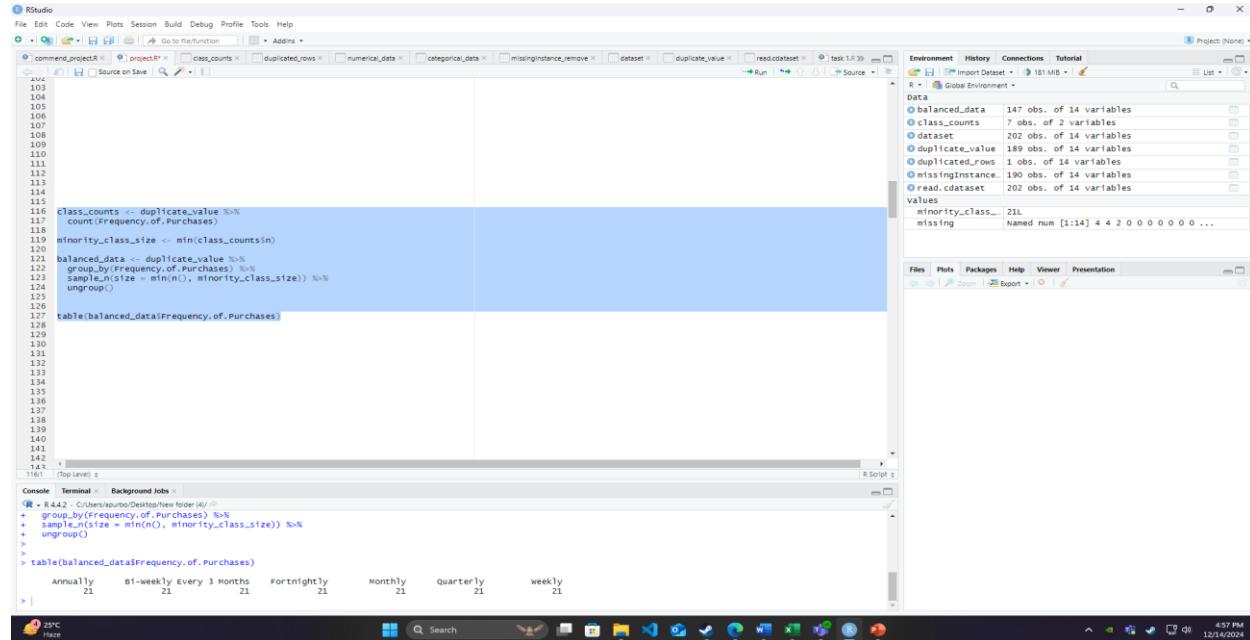
- Data View:** Displays a large dataset table with 189 rows and 14 columns, including columns like Age, Gender, Item.Purchased, Category, Purchase.Amount.USD, Location, Size, Color, Review.Rating, Subscription.Status, Discount.Applied, Promo.Code.Used, Previous.Purchases, and Frequency.
- Code Editor:** Displays the command `duplicate_value <- distinct(missingInstance_remove)`.
- Console:** Shows the command `> which(duplicated(duplicate_value))` being run.
- Environment View:** Shows the variable `dataset` with 202 observations and 14 variables.
- Output View:** Shows the results of the `which` command, indicating no duplicates found.
- System Tray:** Shows the date and time as 12/14/2024 at 4:48 PM.

Convert the imbalanced data set into the balanced data set

We have applied two method for Converting imbalanced to balanced dataset

1. Undersampling
2. Oversampling

Undersampling



```

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
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
599
600
601
602
603
604
605
606
607
608
609
609
610
611
612
613
614
615
616
617
618
619
619
620
621
622
623
624
625
626
627
628
629
629
630
631
632
633
634
635
636
637
638
639
639
640
641
642
643
644
645
646
647
648
649
649
650
651
652
653
654
655
656
657
658
659
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
999
1000
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1097
1098
1099
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1197
1198
1199
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1296
1297
1298
1299
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1396
1397
1398
1399
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1496
1497
1498
1499
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1596
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1696
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2
```

Filter age between 1 to 50

The screenshot shows the RStudio interface with the following components:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menu.
- command_project.R** file open in the script editor.
- Source** tab selected in the script editor.
- Console** pane showing R code execution.
- Terminal** pane showing background jobs.
- Environment** pane listing variables and their types.
- Data** pane showing datasets.
- History** pane showing previous command history.
- Connections** pane.
- Tutorial** pane.
- Project** pane (Project: None).

The script editor contains the following R code:

```
148
149
150
151
152 f1 <- filter(missingInstance_remove, Age > 1 & Age < 50)
153 f2 <- filter(missingInstance_remove, Gender == "Male")
154 f2 <- select(missingInstance_remove,Review.Rating)
155 f3<-select(missingInstance_remove,Review.Rating)
156 f3<-filter(missingInstance_remove,Gender=="Male")
157 f4<-filter(missingInstance_remove,Gender=="Female")
158 f4<-select(missingInstance_remove,Review.Rating)
159 i<-1
152 i (Top Level) 2
```

The console output shows the execution of the R code:

```
R - R4.4.2 C:\Users\user\Desktop\New folder (4).R
> f1 <- filter(missingInstance_remove, Age > 1 & Age < 50)
> f1
```

The environment pane lists the following variables:

- f1: 108 obs. of 14 variables
- f2: 91 obs. of 14 variables
- f3: 190 obs. of 1 variable
- missingInstance_remove: 190 obs. of 14 variables
- read.cdataset: 202 obs. of 14 variables
- task 1: 202 obs. of 14 variables
- values: majority_class_... 36L

The data pane shows the following datasets:

- balanced_data: 288 obs. of 14 variables
- class_counts: 8 obs. of 2 variables
- dataset: 202 obs. of 14 variables
- f1: 108 obs. of 14 variables
- f2: 91 obs. of 14 variables
- f3: 190 obs. of 1 variable
- missingInstance_remove: 190 obs. of 14 variables
- read.cdataset: 202 obs. of 14 variables

The data pane displays the first 37 rows of the dataset:

Age	Gender	Item.Purchased	Category	Purchase.Amount..USD.	Location	Size	color	Review.Rating	subscription.status	discount.applied	
1	19	Male	Sweater	Clothing	64	Maine	L	Maroon	3.1	yes	yes
2	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	3.5	yes	yes
3	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	2.7	yes	yes
4	44	Male	Sneakers	Clothing	20	Wichita	S	White	2.9	yes	yes
5	27	male	shorts	Clothing	34	Louisiana	L	charcoal	3.2	yes	yes
6	30	Male	shorts	Clothing	68	Hawaii	S	olive	4.9	yes	yes
7	23	Male	Sunglasses	Accessories	38	Alabama	S	gray	4.1	yes	yes
8	21	Male	shorts	Clothing	51	Louisiana	M	Black	2.8	yes	yes
9	31	Male	Pants	Clothing	62	North Carolina	M	Charcoal	4.1	yes	yes
10	31	Male	Pants	Clothing	88	Oklahoma	XL	white	4.4	yes	yes
11	31	Male	hoodie	Clothing	25	Texas	M	Silver	3.6	yes	yes
12	38	Male	Jewelry	Accessories	20	Nevada	M	Red	3.6	yes	yes
13	31	Male	Dressesss	Clothing	48	wyoming	S	Black	4.1	yes	yes
14	33	Male	Dress	Clothing	79	west Virginia	L	brown	4.7	yes	yes
15	36	Male	jacket	Clothing	50	Illinois	S	Silver	4.9	yes	yes
16	36	Male	T-shirt	Clothing	91	North Dakota	L	Violet	4.6	yes	yes
17	35	Male	T-shirt	Clothing	69	Illinois	M	Maroon	4.6	yes	yes
18	35	Jeans	Jeans	Clothing	45	Indiana	S	Cyan	2.8	yes	yes
19	35	Male	Dress	Clothing	57	Florida	M	Red	3.7	yes	yes
20	20	Male	coat	Outerwear	100	Tennessee	M	Beige	4.1	yes	yes
21	42	Male	Shirt	Clothing	55	Nevada	M	Orange	2.7	yes	yes
22	49	Male	Blouse	Clothing	28	Maryland	M	Red	3.7	yes	yes
23	41	Male	T-shirt	Clothing	20	Idaho	M	Green	2.6	yes	yes
24	29	Male	Skirt	Clothing	24	Mississippi	S	Peach	3.9	yes	yes
25	47	Male	coat	Outerwear	94	New Mexico	M	Brown	4.2	yes	yes
26	40	Male	jacket	Outerwear	28	South Carolina	M	Lavender	3.0	yes	yes
27	40	Male	Backpacks	Accessories	75	Pennsylvania	M	Green	4.7	yes	yes
28	21	Male	Coat	Outerwear	64	west Virginia	M	white	4.4	yes	yes
29	30	Male	Sneakers	Footwear	21	Mississippi	S	Magenta	3.8	yes	yes
30	33	Male	sunglasses	Accessories	79	Arizona	L	Lavender	2.7	yes	yes
31	33	Male	Blouse	Clothing	33	Pennsylvania	M	Magenta	4.7	yes	yes
32	45	Male	coat	Outerwear	33	Connecticut	L	Gray	4.4	yes	yes
33	48	Male	Belt Accessories	Belt	70	Tennessee	S	Cyan	4.4	yes	yes
34	46	Male	Belt Accessories	Belt	29	Mississippi	M	Magenta	4.2	yes	yes
35	36	Male	Dress	Clothing	48	Virginia	M	Gold	2.9	yes	yes
36	18	Male	Dress	Clothing	26	Georgia	M	olive	2.9	yes	yes
37	48	Male	Pants	Clothing	85	Hawaii	M	teal	2.7	yes	yes

Filter Gender male only

The screenshot shows the RStudio interface with the following components:

- Script Editor:** Displays R code for filtering datasets based on age and gender.
- Data Viewer:** Shows a table of data with columns: Age, Gender, Item, Purchase_Amount, USD, Location, Size, Color, Review_Rating, subscription.status, Discount, Applied_Promo_Code_Used.
- Console:** Displays the results of the R code execution, showing the filtered datasets f1, f2, f3, and f4.

we select only the review.rating column from the dataset

The screenshot shows the RStudio interface with the following details:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menu.
- command.project.R** is the active file.
- Source** tab is selected.
- Code Editor**:
 - Code:

```
148
149
150
152 f1 <- filter(missingInstance_remove, Age > 1 & Age < 30)
153
154 f2 <- filter(missingInstance_remove, Gender == "Male")
155 f2
156 f3<-select(missingInstance_remove,Review.Rating)
157
158 f3
159
160
```
 - Output:

```
R Script 2
```
- Console**:
 - Session path: `R - R 4.2 - C:\Users\aparbo\Desktop\New folder (4)\`
 - Code:

```
> F3<-select(missingInstance_remove,Review.Rating)
> f3
> f3$Rating
```
 - Output:

```
1   Review.Rating
1   3.1
2   3.1
3   3.1
4   3.5
5   2.7
6   2.1
7   3.2
8   3.2
9   4.8
10  4.1
11  4.1
12  4.9
13  4.5
14  2.8
15  4.1
16  4.1
17  3.3
18  3.3
19  2.8
20  3.3
21  2.8
22  4.1
23  3.2
24  4.4
25  3.6
26  3.6
27  3.6
28  5.0
29  4.4
30  4.1
31  4.7
32  4.7
33  4.9
34  3.3
35  4.6
36  4.0
37  4.6
38  2.8
39  3.7
40  4.2
41  4.6
42  4.5
43  4.1
```
- Environment** pane:
 - Data:
 - balanced_data: 288 obs. of 14 variables
 - class_counts: 8 obs. of 2 variables
 - dataset: 202 obs. of 14 variables
 - f1: 108 obs. of 14 variables
 - f2: 91 obs. of 14 variables
 - f3: 190 obs. of 1 variable
 - missingInstance_remove: 190 obs. of 14 variables
 - read_cdataset: 202 obs. of 14 variables
 - values:
 - majority_class: 36L
- Project** pane: `None`.
- Bottom Bar**: `25°C Haze`, `Search`, `Task View`, `File Explorer`, `Terminal`, `Plots`, `Packages`, `Help`, `Viewer`, `Presentation`.
- Status Bar**: `5:12 PM`, `12/14/2023`.

Convert attributes from numeric to categorical

We categorize numerical columns into groups. Age is categorized as Young =0–23, Adult =24–38, Middle-aged =39–58, and Senior =+59. Purchase.Amount..USD. is grouped into Low spender =0–49, Medium spender=50–99, and High spender=+100. Review.Rating is labeled as Poor= 0–2.9, Average =3–3.9, Good =4–4.9, and Excellent= 5. Previous.Purchases is grouped into Few Purchases=0–9, Moderate Purchases= 10–49, and Frequent Purchases =+50.

The screenshot shows the RStudio interface with the following details:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menu items.
- Above the main workspace, there are tabs for "command.project.R" and "project.R".
- The main workspace displays a data frame named "numerical_Categorical" with 189 rows and 5 columns. The columns are "Age", "Purchase.Amount.USD.", "Review.Rating", "Previous.Purchases", and "Spending.Type". The data includes rows like "1 Middle-aged Medium spender Average Moderate Purchases" and "19 Young Medium spender Good Few Purchases".
- The status bar at the bottom left shows "2 BDT/RUB -4.6%".
- The status bar at the bottom right shows "5:31 PM 12/14/2024".
- Global Environment** pane on the right side:
 - Shows objects: "balanced_data", "class_counts", "dataset", "duplicate_value", "f1", "f2", "f3", "missinginstance_remove", "numerical_catag", "read.cdataset", "values".
 - Object descriptions:
 - "balanced_data": 288 obs. of 14 variables
 - "class_counts": 8 obs. of 2 variables
 - "dataset": 202 obs. of 14 variables
 - "duplicate_value": 189 obs. of 14 variables
 - "f1": 108 obs. of 14 variables
 - "f2": 91 obs. of 14 variables
 - "f3": 190 obs. of 1 variable
 - "missinginstance_remove": 190 obs. of 14 variables
 - "numerical_catag": 189 obs. of 4 variables
 - "read.cdataset": 202 obs. of 14 variables
 - Object types: "values", "column_check".
 - Buttons: "Named", "log1 [1:14]", "TRUE FALSE FALSE FALSE TRUE".
 - Navigation buttons: "Files", "Plots", "Packages", "Help", "Viewer", "Presentation".

convert attributes from categorical to numeric

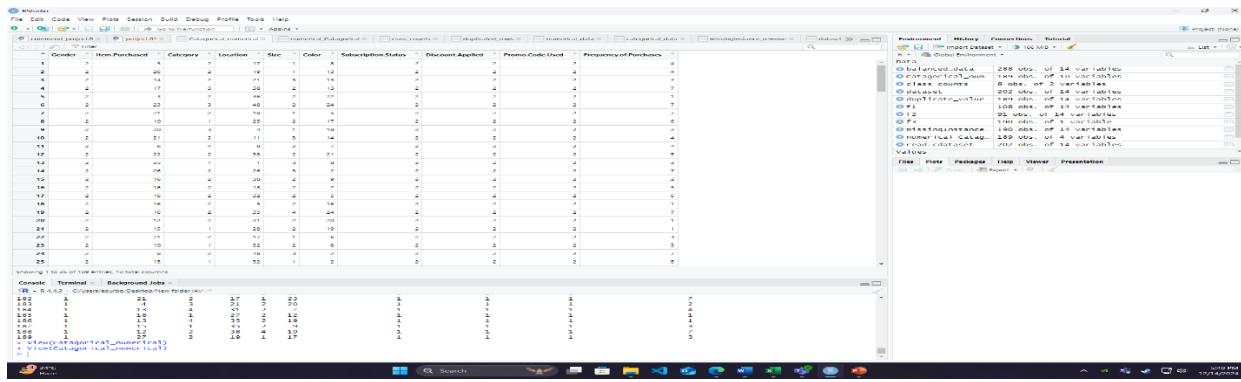
We convert categorical columns into numeric values. Gender, Item.Purchased, Category, Location, Size, Color, Subscription.Status, Discount.Applied, Promo.Code.Used, and Frequency.of.Purchases are all converted to numeric codes where each unique category is replaced with a unique numeric value.

The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a Go to file/function button. The top right corner shows a Project (None) status. The left sidebar has tabs for command_project.R, project.R, numerical_Categorical, class_counts, duplicated_rows, numerical_data, categorical_data, missinginstance_remove, dataset, duplicate_value, and Source. The main area contains the following R code:

```
187
188 #ategorical_numerical <- duplicate_value[, column_check]
189
190 Categorical_numerical$Gender <- as.numeric(factor(Categorical_numerical$Gender))
191
192 Categorical_numerical$item.Purchased <- as.numeric(factor(Categorical_numerical$item.Purchased))
193
194 Categorical_numerical$category <- as.numeric(factor(Categorical_numerical$category))
195
196 Categorical_numerical$location <- as.numeric(factor(Categorical_numerical$location))
197
198 Categorical_numerical$size <- as.numeric(factor(Categorical_numerical$size))
199
200 Categorical_numerical$color <- as.numeric(factor(Categorical_numerical$color))
201
202 Categorical_numerical$subscription.status <- as.numeric(factor(Categorical_numerical$subscription.status))
203
204 Categorical_numerical$discount.Applied <- as.numeric(factor(Categorical_numerical$discount.Applied))
205
206 Categorical_numerical$promo.code.used <- as.numeric(factor(Categorical_numerical$promo.code.used))
207
208 Categorical_numerical$Frequency.of.Purchases <- as.numeric(factor(Categorical_numerical$Frequency.of.Purchases))
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226 Categorical_numerical
```

The bottom left shows the R console with the command `> Categorical_numerical` and the resulting output:

	Gender	Item.Purchased	Category	Location	size	color	subscription.status	Discount.Applied	Promo.Code.used	Frequency.of.Purchases
1	2	3	2	17	1	8	2	2	2	4
2	2	26	2	2	1	13	2	2	2	4
3	2	14	2	21	3	13	2	2	2	7
4	2	17	3	38	2	13	2	2	2	7
5	2	3	2	38	2	22	2	2	2	1
6	2	23	3	2	2	24	2	2	2	7
7	2	21	2	18	1	5	2	2	2	7
8	2	10	1	25	2	17	2	2	2	6
9	2	20	3	4	1	18	2	2	2	2



Outliers

We select specific attributes from the dataset, identified by column_check, and store them in numerical_Catagorical_normalization. Then, we generate a summary of the dataset, which provides statistical details about the data.

```

File Edit Code View Plots Session Build Debug Profile Tools Help
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R > R 4.4.1 - /r
C:\Users\user\OneDrive\Documents\GitHub\ProjectR>
numerical_Catagorical_normalization <- duplicate_value[,column_check]
summary(numerical_Catagorical_normalization)
Q1 <- quantile(numerical_Catagorical_normalization$Age, 0.25, na.rm = TRUE)
Q3 <- quantile(numerical_Catagorical_normalization$Age, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
outliers <- numerical_Catagorical_normalization[normals$Age < lower_bound | numerical_Catagorical_normalization[normals$Age > upper_bound], ]
numerical_Catagorical_normalization <- numerical_Catagorical_normalization[-normals, ]
summary(numerical_Catagorical_normalization)
Age Purchase.Amount..USD. Review.Rating Previous.Purchases
Min. : 18.00 Min. : 20.00 Min. :2,500 Min. : 1.00
1st Qu.: 31.00 1st Qu.: 40.00 1st Qu.:3,100 1st Qu.:14.00
Median : 45.00 Median : 62.00 Median :3,800 Median :27.00
Mean : 45.44 Mean : 61.18 Mean :3,772 Mean :26.73
3rd Qu.: 56.00 3rd Qu.: 82.00 3rd Qu.:4,400 3rd Qu.:40.00
Max. :256.00 Max. :100.00 Max. :5,000 Max. :50.00

```

We calculate the first (Q1) and third quartiles (Q3) of the age column in categorical normalization. Then we compute the Interquartile Range (IQR) by subtracting Q1 from Q3. The lower and upper bounds are determined using 1.5 times the IQR. At last we identify the outliers in the Age attribute that fall outside these bounds and store them in the outliers variable.

The screenshot shows the RStudio interface with the following details:

- File menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project pane:** command_project.R, project.R, Categorical_numerical, numerical_Categorical, class_counts, duplicated_rows, numerical_data, categorical_data, missinginstance_remove, dataset.
- Environment pane:** Global Environment, numerical_Catag., outliers_remove, read_cddataset.
- Console pane:**

```
R> R.4.4.2 - C:\Users\apropo\Desktop\New folder (4)\>
> >
> numerical_categorical_normalization <- duplicate_value[,column_check]
> summary(numerical_categorical_normalization)
> Q1 <- quantile(numerical_categorical_normalization$Age, 0.25, na.rm = TRUE)
> Q3 <- quantile(numerical_categorical_normalization$Age, 0.75, na.rm = TRUE)
> IQR <- Q3 - Q1
> 
> lower_bound <- Q1 - 1.5 * IQR
> upper_bound <- Q3 + 1.5 * IQR
> 
> outliers <- numerical_categorical_normalization$Age[numerical_categorical_normalization$Age < lower_bound | numerical_categorical_normalization$Age > upper_bound]
> outliers
[1] 256 235
>
```
- Plots pane:** Not visible.
- Packages pane:** Not visible.
- Help pane:** Not visible.
- Presentation pane:** Not visible.
- Bottom status bar:** 12:31 PM, 12/14/2024.

We filter the Categorical_normalization_normalization dataset to remove the outliers from the Age column by keeping only the intence where Age is between the lower bound and upper bound. The resulting dataset outliers remove is then summarized to provide statistical details.

The screenshot shows the RStudio interface with the following details:

- File menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project pane:** command_project.R, project.R, Categorical_numerical, numerical_Categorical, class_counts, duplicated_rows, numerical_data, categorical_data, missinginstance_remove, dataset.
- Environment pane:** Global Environment, numerical_Catag., outliers_remove, read_cddataset.
- Console pane:**

```
R> R.4.4.2 - C:\Users\apropo\Desktop\New folder (4)\>
> >
> outliers <- numerical_categorical_normalization$Age[numerical_categorical_normalization$Age < lower_bound | numerical_categorical_normalization$Age > upper_bound]
> outliers
[1] 256 235
> outliers_remove <- numerical_categorical_normalization %>% filter(Age == lower_bound & Age == upper_bound)
> 
> summary(outliers_remove)
Age   Purchase.Amount   USD. review.Rating   Previous.Purchases
Min. :38.0   Min. :40.00   Min. :2.500   Min. : 1.00
1st Qu.:31.0  1st Qu.:40.00   1st Qu.:3.100  1st Qu.:14.50
Median :44.0  Median :62.00   Median :3.800  Median :27.00
Mean   :44.9  Mean   :62.00   Mean   :3.844  Mean   :27.9
3rd Qu.:55.5  3rd Qu.:82.50   3rd Qu.:4.400  3rd Qu.:40.00
Max. :70.0   Max. :100.00  Max. :5.000  Max. :50.00
>
```
- Plots pane:** Not visible.
- Packages pane:** Not visible.
- Help pane:** Not visible.
- Presentation pane:** Not visible.
- Bottom status bar:** 6:14 PM, 12/14/2024.

normalization method

Normalizes the Review.Rating column to a 0–1 scale adds it as a new column (Rating_Normalized) and then removes the original Review.Rating column from the dataset.