SHETH L.U.J.AND SIR M.V. COLLEGE
PRACTICAL NO.13
Data Analysis with SAS / SPSS /R

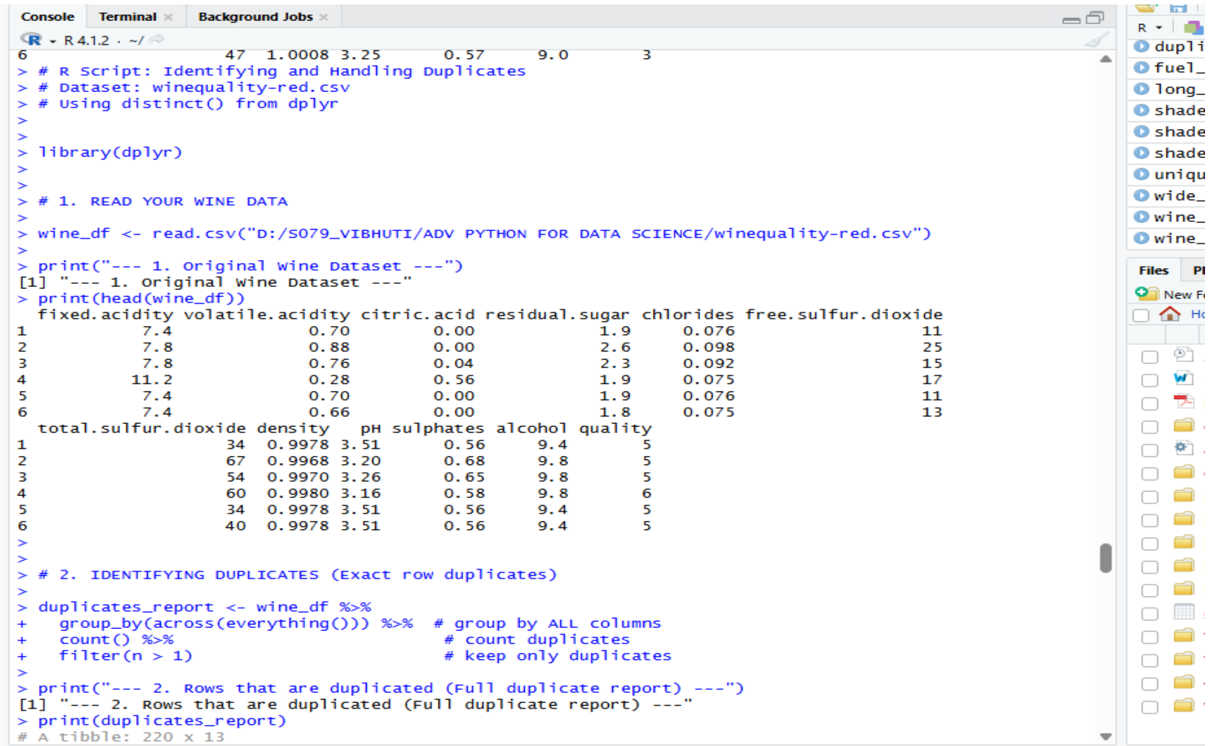
AIM:- Identifying and handling duplicates using distinct() (R studio).

INPUT:-

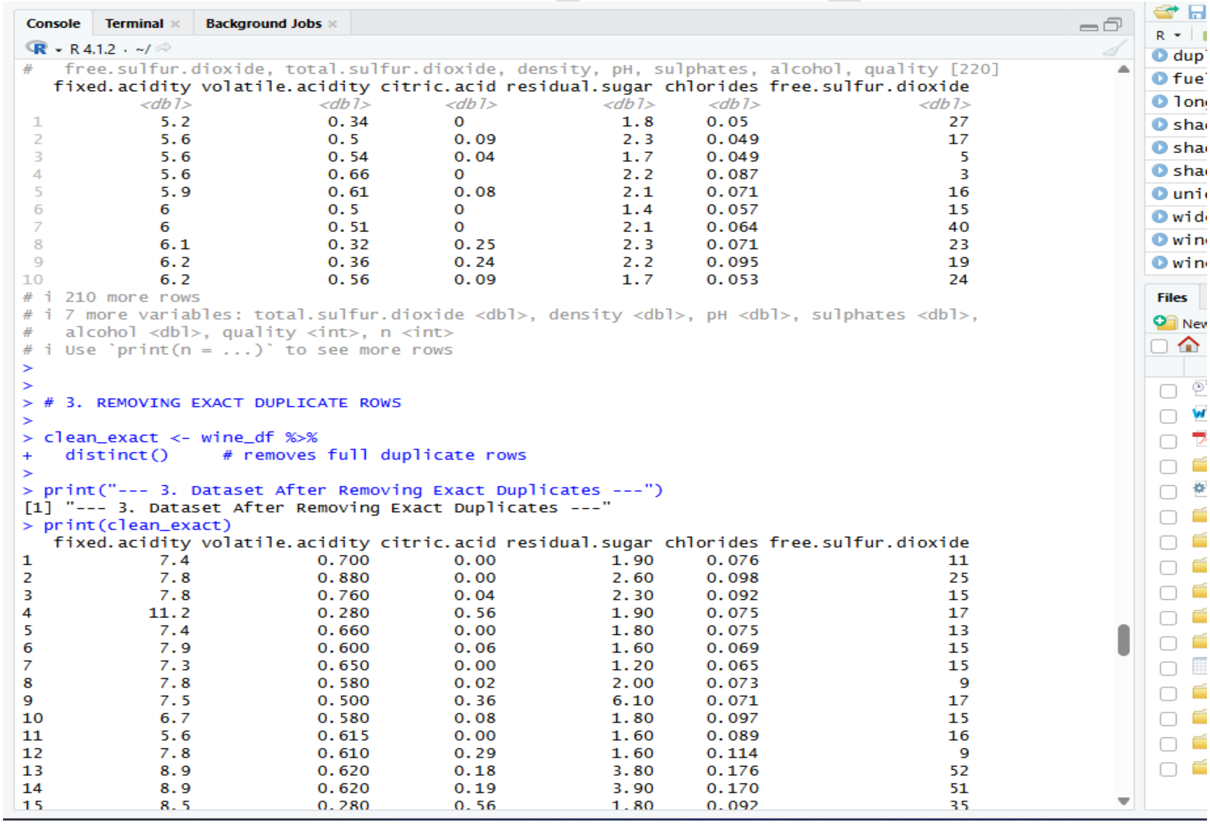
```
# R Script: Identifying and Handling Duplicates

# Dataset: winequality-red.csv

# Using distinct() from dplyr


library(dplyr)

# 1. READ YOUR WINE DATA

wine_df <- read.csv("D:/S079_VIBHUTI/ADV PYTHON FOR DATA
SCIENCE/winequality-red.csv")


print("--- 1. Original Wine Dataset ---")

print(head(wine_df))



# 2. IDENTIFYING DUPLICATES (Exact row duplicates)


duplicates_report <- wine_df %>%
  group_by(across(everything())) %>%  # group by ALL columns
  count() %>%                          # count duplicates
  filter(n > 1)                      # keep only duplicates


print("--- 2. Rows that are duplicated (Full duplicate report) ---")

print(duplicates_report)
```

```
# 3. REMOVING EXACT DUPLICATE ROWS

clean_exact <- wine_df %>%
  distinct()     # removes full duplicate rows

print("--- 3. Dataset After Removing Exact Duplicates ---")
print(clean_exact)


# 4. HANDLING DUPLICATES BASED ON ONE COLUMN (Example: quality)

# Scenario: Keep ONLY ONE ROW per quality level
# This is like your 'unique customers' example.

unique_quality <- wine_df %>%
  distinct(quality, .keep_all = TRUE)

print("--- 4. Unique Quality Values (Only first appearance kept) ---")
print(unique_quality)
```

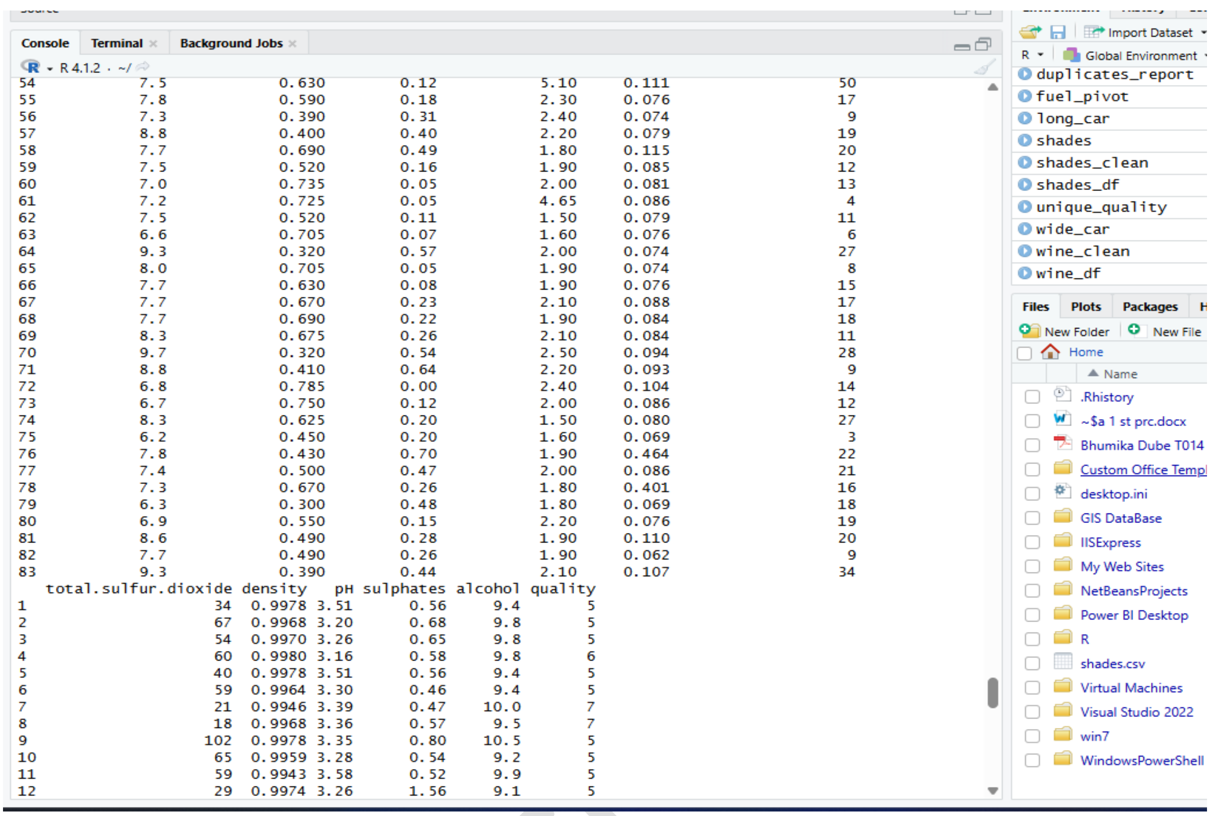
OUTPUT:-



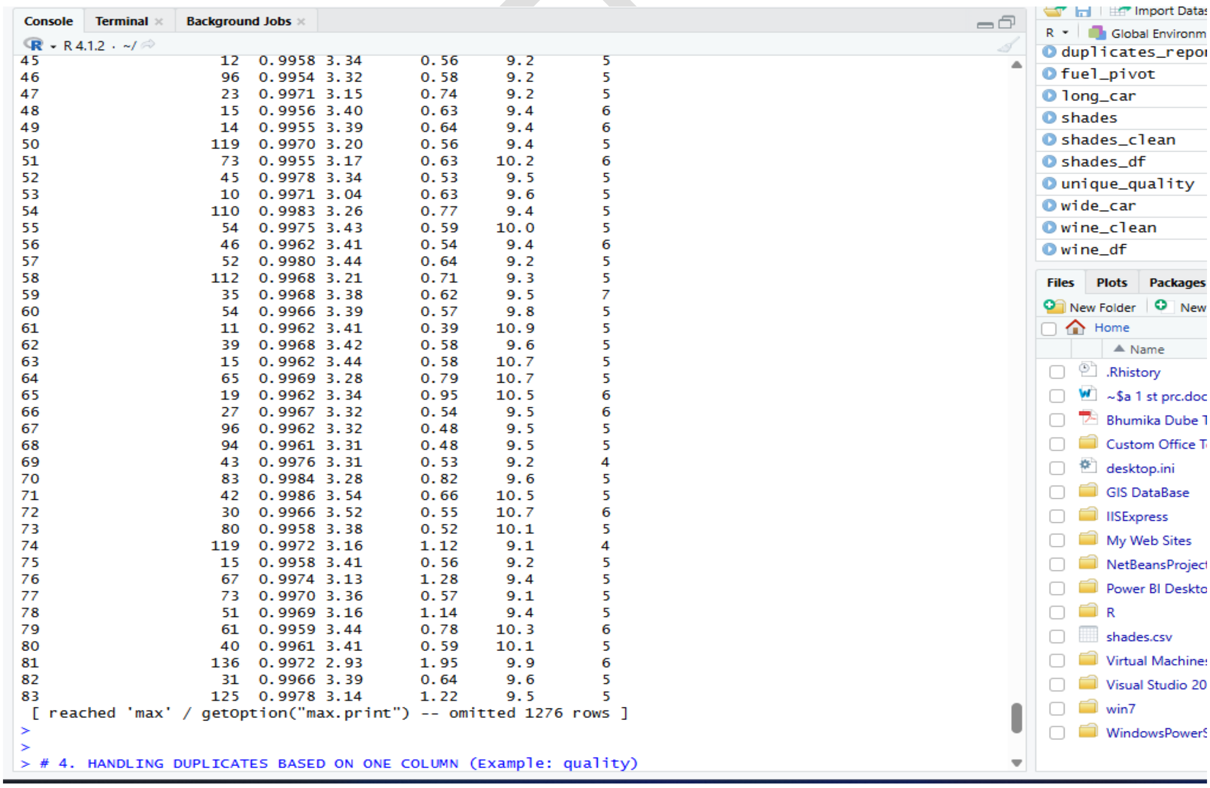
```
6                      47  1.0008 3.25      0.57     9.0       3
> # R Script: Identifying and Handling Duplicates
> # Dataset: winequality-red.csv
> # Using distinct() from dplyr
>
>
> library(dplyr)
>
>
> # 1. READ YOUR WINE DATA
>
> wine_df <- read.csv("D:/S079_VIBHUTI/ADV PYTHON FOR DATA SCIENCE/winequality-red.csv")
>
> print("--- 1. Original Wine Dataset ---")
[1] "--- 1. Original Wine Dataset ---"
> print(head(wine_df))
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1           7.4             0.70        0.00            1.9     0.076                  11
2           7.8             0.88        0.00            2.6     0.098                  25
3           7.8             0.76        0.04            2.3     0.092                  15
4          11.2             0.28        0.56            1.9     0.075                  17
5           7.4             0.70        0.00            1.9     0.076                  11
6           7.4             0.66        0.00            1.8     0.075                  13
  total.sulfur.dioxide density   pH sulphates alcohol quality
1                   34  0.9978 3.51      0.56     9.4       5
2                   67  0.9968 3.20      0.68     9.8       5
3                   54  0.9970 3.26      0.65     9.8       5
4                   60  0.9980 3.16      0.58     9.8       6
5                   34  0.9978 3.51      0.56     9.4       5
6                   40  0.9978 3.51      0.56     9.4       5
>
>
> # 2. IDENTIFYING DUPLICATES (Exact row duplicates)
>
> duplicates_report <- wine_df %>%
+   group_by(across(everything())) %>%  # group by ALL columns
+   count() %>%                          # count duplicates
+   filter(n > 1)                        # keep only duplicates
>
> print("--- 2. Rows that are duplicated (Full duplicate report) ---")
[1] "--- 2. Rows that are duplicated (Full duplicate report) ---"
> print(duplicates_report)
# A tibble: 220 x 13
```



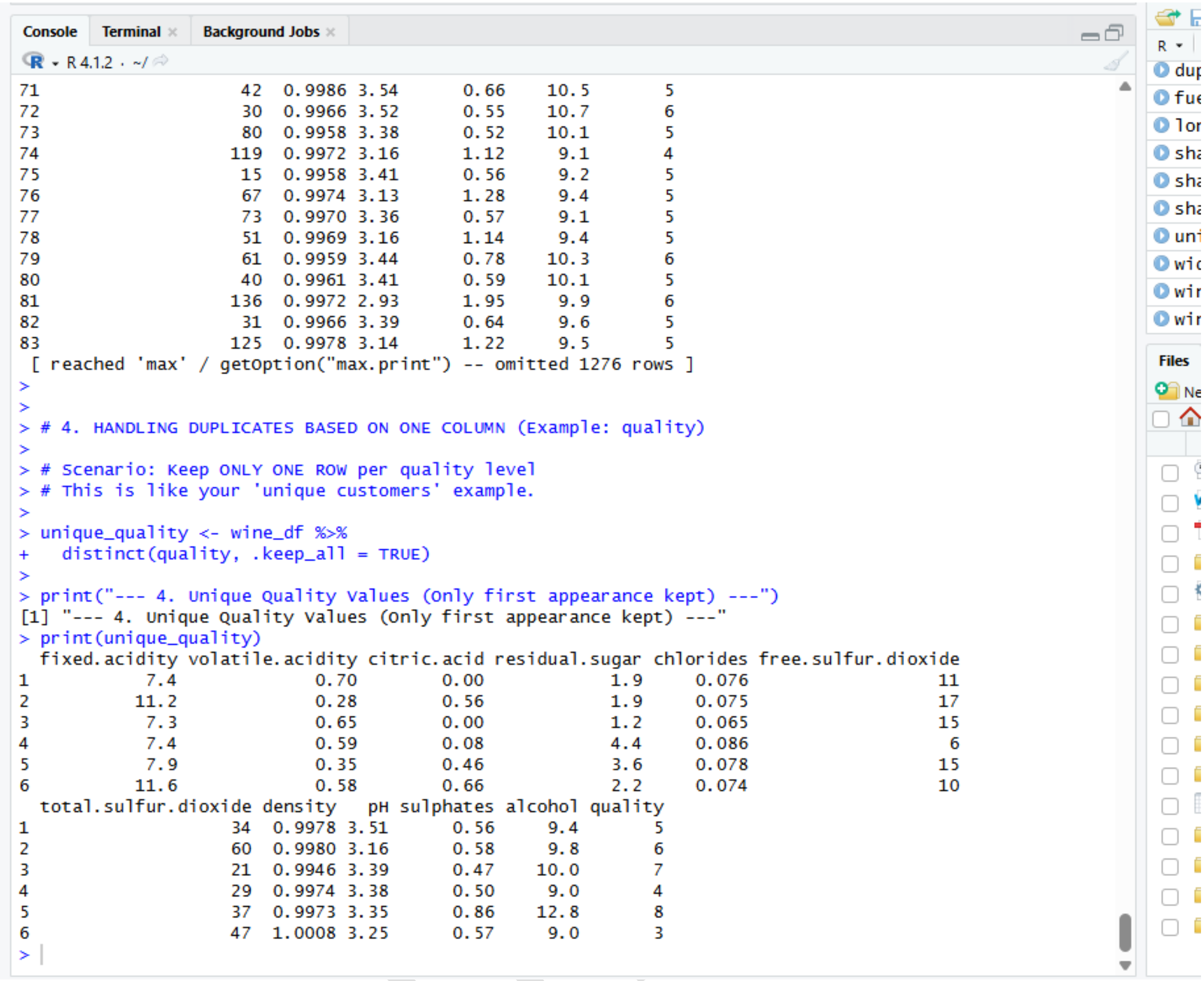
```
#   free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality [220]
   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>               <dbl>
 1           5.2             0.34        0               1.8     0.05                   27
 2           5.6             0.5         0.09            2.3     0.049                  17
 3           5.6             0.54        0.04            1.7     0.049                   5
 4           5.6             0.66        0               2.2     0.087                   3
 5           5.9             0.61        0.08            2.1     0.071                  16
 6           6               0.5         0               1.4     0.057                  15
 7           6               0.51        0               2.1     0.064                  40
 8           6.1             0.32        0.25            2.3     0.071                  23
 9           6.2             0.36        0.24            2.2     0.095                  19
10           6.2             0.56        0.09            1.7     0.053                  24
# i 210 more rows
# i 7 more variables: total.sulfur.dioxide <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
#   alcohol <dbl>, quality <int>, n <int>
# i Use `print(n = ...)` to see more rows
>
>
> # 3. REMOVING EXACT DUPLICATE ROWS
>
> clean_exact <- wine_df %>%
+   distinct()      # removes full duplicate rows
>
> print("--- 3. Dataset After Removing Exact Duplicates ---")
[1] "--- 3. Dataset After Removing Exact Duplicates ---"
> print(clean_exact)
   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1            7.4            0.700        0.00           1.90     0.076                  11
2            7.8            0.880        0.00           2.60     0.098                  25
3            7.8            0.760        0.04           2.30     0.092                  15
4           11.2            0.280        0.56           1.90     0.075                  17
5            7.4            0.660        0.00           1.80     0.075                  13
6            7.9            0.600        0.06           1.60     0.069                  15
7            7.3            0.650        0.00           1.20     0.065                  15
8            7.8            0.580        0.02           2.00     0.073                   9
9            7.5            0.500        0.36           6.10     0.071                  17
10           6.7            0.580        0.08           1.80     0.097                  15
11           5.6            0.615        0.00           1.60     0.089                  16
12           7.8            0.610        0.29           1.60     0.114                   9
13           8.9            0.620        0.18           3.80     0.176                  52
14           8.9            0.620        0.19           3.90     0.170                  51
15           8.5            0.280        0.56           1.80     0.092                  35
```



```
54           7.5            0.630        0.12           5.10     0.111                  50
55           7.8            0.590        0.18           2.30     0.076                  17
56           7.3            0.390        0.31           2.40     0.074                   9
57           8.8            0.400        0.40           2.20     0.079                  19
58           7.7            0.690        0.49           1.80     0.115                  20
59           7.5            0.520        0.16           1.90     0.085                  12
60           7.0            0.735        0.05           2.00     0.081                  13
61           7.2            0.725        0.05           4.65     0.086                   4
62           7.5            0.520        0.11           1.50     0.079                  11
63           6.6            0.705        0.07           1.60     0.076                   6
64           9.3            0.320        0.57           2.00     0.074                  27
65           8.0            0.705        0.05           1.90     0.074                   8
66           7.7            0.630        0.08           1.90     0.076                  15
67           7.7            0.670        0.23           2.10     0.088                  17
68           7.7            0.690        0.22           1.90     0.084                  18
69           8.3            0.675        0.26           2.10     0.084                  11
70           9.7            0.320        0.54           2.50     0.094                  28
71           8.8            0.410        0.64           2.20     0.093                   9
72           6.8            0.785        0.00           2.40     0.104                  14
73           6.7            0.750        0.12           2.00     0.086                  12
74           8.3            0.625        0.20           1.50     0.080                  27
75           6.2            0.450        0.20           1.60     0.069                   3
76           7.8            0.430        0.70           1.90     0.464                  22
77           7.4            0.500        0.47           2.00     0.086                  21
78           7.3            0.670        0.26           1.80     0.401                  16
79           6.3            0.300        0.48           1.80     0.069                  18
80           6.9            0.550        0.15           2.20     0.076                  19
81           8.6            0.490        0.28           1.90     0.110                  20
82           7.7            0.490        0.26           1.90     0.062                   9
83           9.3            0.390        0.44           2.10     0.107                  34
   total.sulfur.dioxide density   pH sulphates alcohol quality
1                    34  0.9978 3.51      0.56     9.4       5
2                    67  0.9968 3.20      0.68     9.8       5
3                    54  0.9970 3.26      0.65     9.8       5
4                    60  0.9980 3.16      0.58     9.8       6
5                    40  0.9978 3.51      0.56     9.4       5
6                    59  0.9964 3.30      0.46     9.4       5
7                    21  0.9946 3.39      0.47    10.0       7
8                    18  0.9968 3.36      0.57     9.5       7
9                   102  0.9978 3.35      0.80    10.5       5
10                   65  0.9959 3.28      0.54     9.2       5
11                   59  0.9943 3.58      0.52     9.9       5
12                   29  0.9974 3.26      1.56     9.1       5
```



```
45                   12  0.9958 3.34      0.56     9.2       5
46                   96  0.9954 3.32      0.58     9.2       5
47                   23  0.9971 3.15      0.74     9.2       5
48                   15  0.9956 3.40      0.63     9.4       6
49                   14  0.9955 3.39      0.64     9.4       6
50                  119  0.9970 3.20      0.56     9.4       5
51                   73  0.9955 3.17      0.63    10.2       6
52                   45  0.9978 3.34      0.53     9.5       5
53                   10  0.9971 3.04      0.63     9.6       5
54                  110  0.9983 3.26      0.77     9.4       5
55                   54  0.9975 3.43      0.59    10.0       5
56                   46  0.9962 3.41      0.54     9.4       6
57                   52  0.9980 3.44      0.64     9.2       5
58                  112  0.9968 3.21      0.71     9.3       5
59                   35  0.9968 3.38      0.62     9.5       7
60                   54  0.9966 3.39      0.57     9.8       5
61                   11  0.9962 3.41      0.39    10.9       5
62                   39  0.9968 3.42      0.58     9.6       5
63                   15  0.9962 3.44      0.58    10.7       5
64                   65  0.9969 3.28      0.79    10.7       5
65                   19  0.9962 3.34      0.95    10.5       6
66                   27  0.9967 3.32      0.54     9.5       6
67                   96  0.9962 3.32      0.48     9.5       5
68                   94  0.9961 3.31      0.48     9.5       5
69                   43  0.9976 3.31      0.53     9.2       4
70                   83  0.9984 3.28      0.82     9.6       5
71                   42  0.9986 3.54      0.66    10.5       5
72                   30  0.9966 3.52      0.55    10.7       6
73                   80  0.9958 3.38      0.52    10.1       5
74                  119  0.9972 3.16      1.12     9.1       4
75                   15  0.9958 3.41      0.56     9.2       5
76                   67  0.9974 3.13      1.28     9.4       5
77                   73  0.9970 3.36      0.57     9.1       5
78                   51  0.9969 3.16      1.14     9.4       5
79                   61  0.9959 3.44      0.78    10.3       6
80                   40  0.9961 3.41      0.59    10.1       5
81                  136  0.9972 2.93      1.95     9.9       6
82                   31  0.9966 3.39      0.64     9.6       5
83                  125  0.9978 3.14      1.22     9.5       5
 [ reached 'max' / getOption("max.print") -- omitted 1276 rows ]
>
>
> # 4. HANDLING DUPLICATES BASED ON ONE COLUMN (Example: quality)
```



```
71                   42  0.9986 3.54      0.66    10.5       5
72                   30  0.9966 3.52      0.55    10.7       6
73                   80  0.9958 3.38      0.52    10.1       5
74                  119  0.9972 3.16      1.12     9.1       4
75                   15  0.9958 3.41      0.56     9.2       5
76                   67  0.9974 3.13      1.28     9.4       5
77                   73  0.9970 3.36      0.57     9.1       5
78                   51  0.9969 3.16      1.14     9.4       5
79                   61  0.9959 3.44      0.78    10.3       6
80                   40  0.9961 3.41      0.59    10.1       5
81                  136  0.9972 2.93      1.95     9.9       6
82                   31  0.9966 3.39      0.64     9.6       5
83                  125  0.9978 3.14      1.22     9.5       5
 [ reached 'max' / getOption("max.print") -- omitted 1276 rows ]
>
>
> # 4. HANDLING DUPLICATES BASED ON ONE COLUMN (Example: quality)
>
> # Scenario: Keep ONLY ONE ROW per quality level
> # This is like your 'unique customers' example.
>
> unique_quality <- wine_df %>%
+   distinct(quality, .keep_all = TRUE)
>
> print("--- 4. Unique Quality Values (Only first appearance kept) ---")
[1] "--- 4. Unique Quality Values (Only first appearance kept) ---"
> print(unique_quality)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide
1           7.4             0.70        0.00            1.9     0.076                  11
2          11.2             0.28        0.56            1.9     0.075                  17
3           7.3             0.65        0.00            1.2     0.065                  15
4           7.4             0.59        0.08            4.4     0.086                   6
5           7.9             0.35        0.46            3.6     0.078                  15
6          11.6             0.58        0.66            2.2     0.074                  10
  total.sulfur.dioxide density   pH sulphates alcohol quality
1                   34  0.9978 3.51      0.56     9.4       5
2                   60  0.9980 3.16      0.58     9.8       6
3                   21  0.9946 3.39      0.47    10.0       7
4                   29  0.9974 3.38      0.50     9.0       4
5                   37  0.9973 3.35      0.86    12.8       8
6                   47  1.0008 3.25      0.57     9.0       3
> |
```