Analysis on red wine dataset

Made by:
Dharini Patel A201
Khushi Patil A216
Vishwasinh Suratia A231

CONTENTS

01

INTRODUCTION

Introduction to dataset

02

R STUDIO

Loading dataset in R studio

03

ANALYSIS

Analysis of the dataset

04

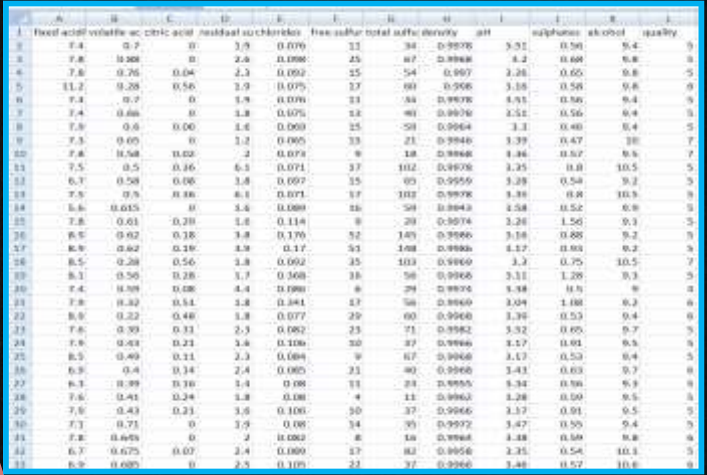
CONCLUSION

01

INTRODUCTION

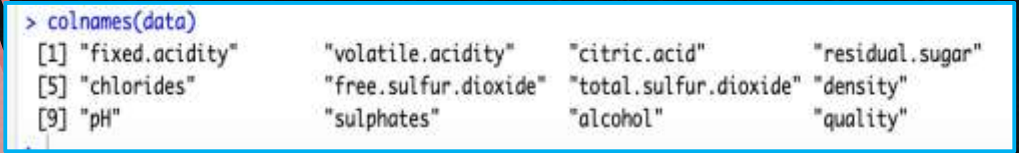
Introduction to Dataset

Going through drinks, which is way more popular in this era gave curiosity about how much level of harm it can cause or not and by going through a ranked wine data made it possible to come to a conclusion



	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidit	volatile ac	citric acid	residual su	chlorides	free sulfur	total sulfu	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
18	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
19	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
20	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
21	7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
22	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
23	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
24	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
25	8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
26	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
27	6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
28	7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5
29	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
30	7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5
31	7.8	0.645	0	2	0.082	8	16	0.9964	3.38	0.59	9.8	6
32	6.7	0.675	0.07	2.4	0.089	17	82	0.9958	3.35	0.54	10.1	5
33	6.9	0.685	0	2.5	0.105	22	37	0.9966	3.46	0.57	10.6	6

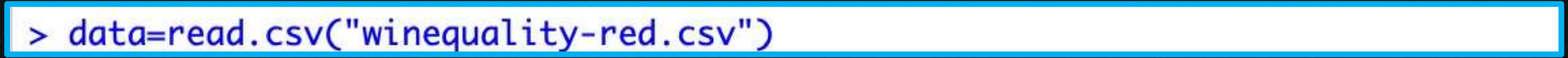
Link: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=downloadhttps://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?resource=download



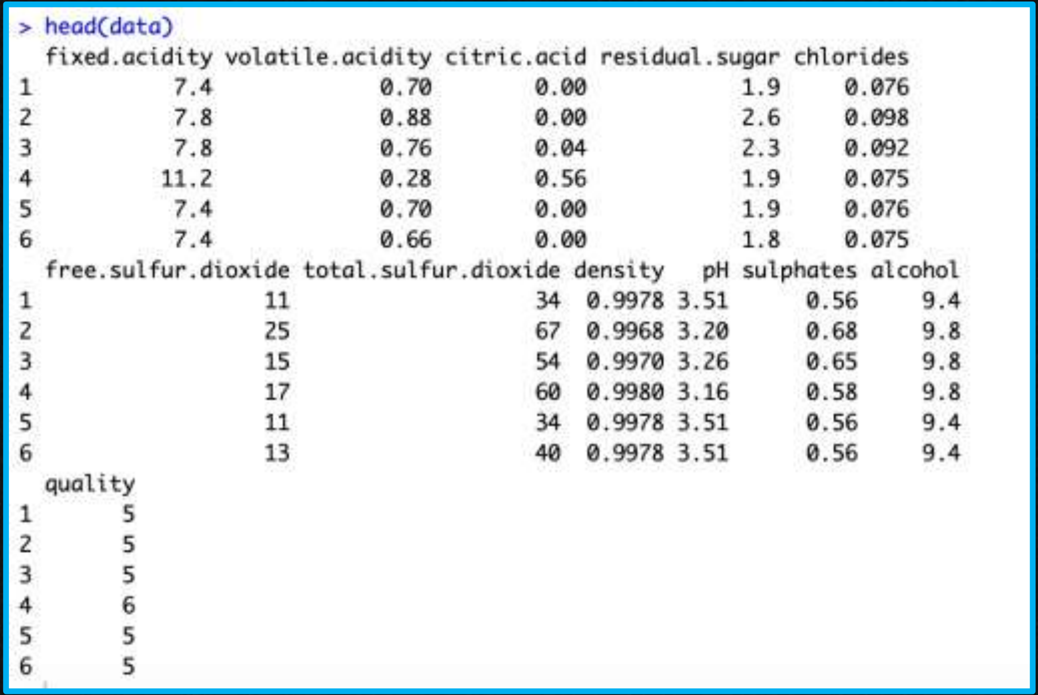
```
> colnames(data)
 [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"          "residual.sugar"
 [5] "chlorides"            "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
 [9] "pH"                   "sulphates"            "alcohol"              "quality"
```

02

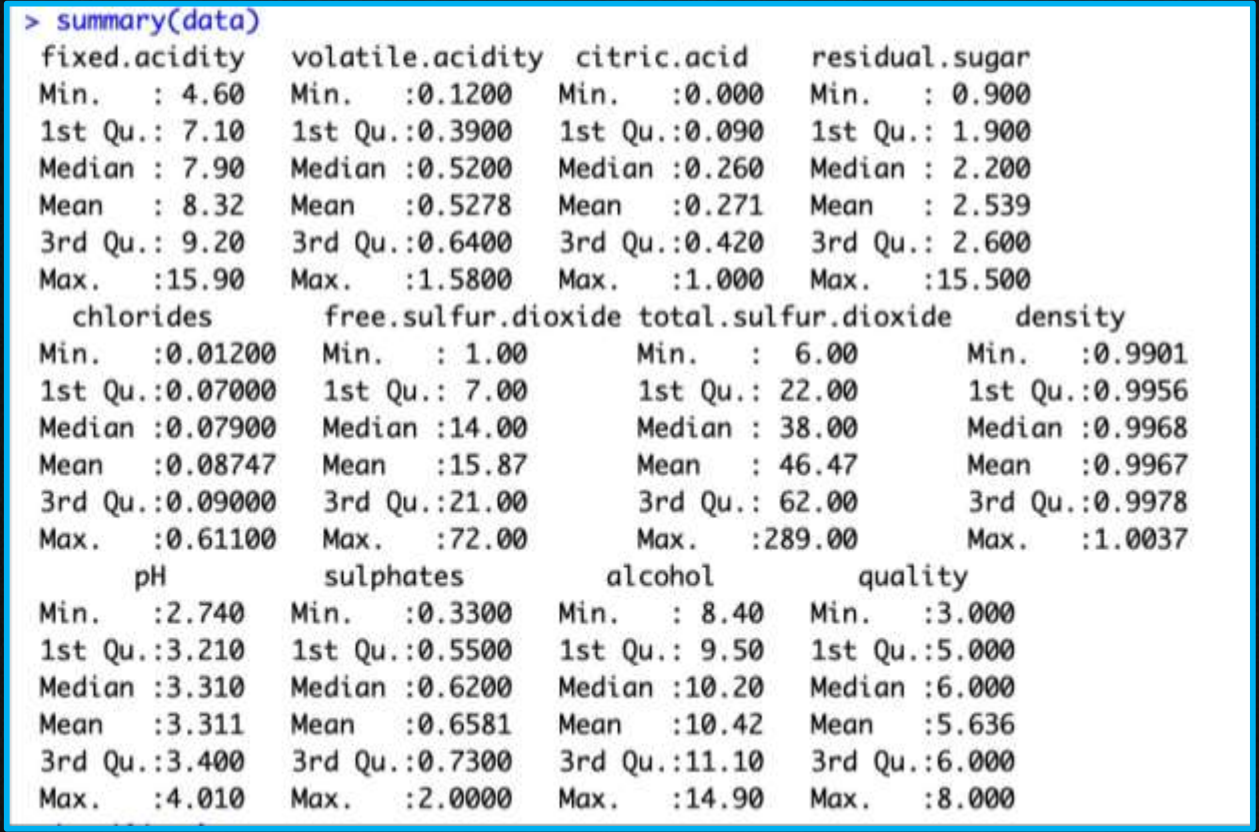
DATASET IN R STUDIO



```
> data=read.csv("winequality-red.csv")
```



```
> head(data)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1           7.4             0.70        0.00            1.9     0.076
2           7.8             0.88        0.00            2.6     0.098
3           7.8             0.76        0.04            2.3     0.092
4          11.2             0.28        0.56            1.9     0.075
5           7.4             0.70        0.00            1.9     0.076
6           7.4             0.66        0.00            1.8     0.075
  free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
1                  11                   34  0.9978 3.51      0.56     9.4
2                  25                   67  0.9968 3.20      0.68     9.8
3                  15                   54  0.9970 3.26      0.65     9.8
4                  17                   60  0.9980 3.16      0.58     9.8
5                  11                   34  0.9978 3.51      0.56     9.4
6                  13                   40  0.9978 3.51      0.56     9.4
  quality
1       5
2       5
3       5
4       6
5       5
6       5
```



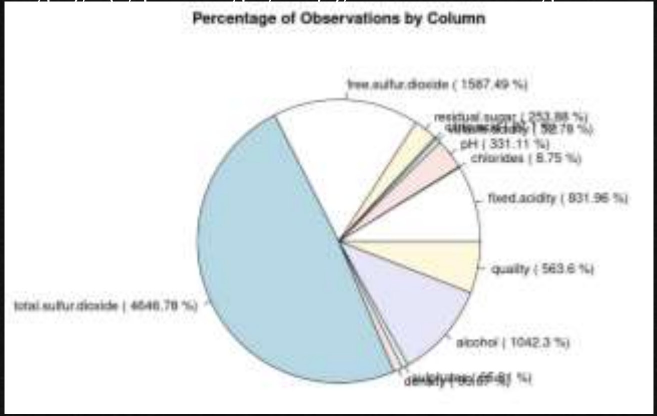
```
> summary(data)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
       pH          sulphates         alcohol         quality     
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
 Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
```

03

Analysis of the Dataset

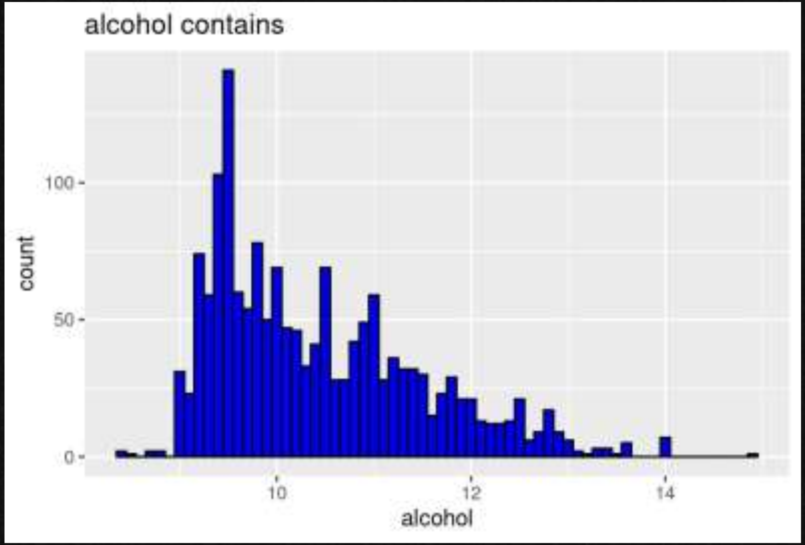
Physicochemical OBS: using pie chart

```
# select the columns of interest
wine_sub <- select (wine, fixed.acidity, chlorides, pH, volatile.acidity,citric.acid, residual.sugar, free.sulfur.dioxide,
total.sulfur.dioxide, density, sulphates, alcohol, quality)

# calculate the total number of observations
total_obs <- nrow(wine_sub)

# calculate the percentage of observations for each column
percentages <- round(colSums(wine_sub) / total_obs * 100, 2)

# create the pie chart
pie (percentages, labels = paste(names(percentages), "(", percentages, "%)"), main = "Percentage of Observations by Column")
```



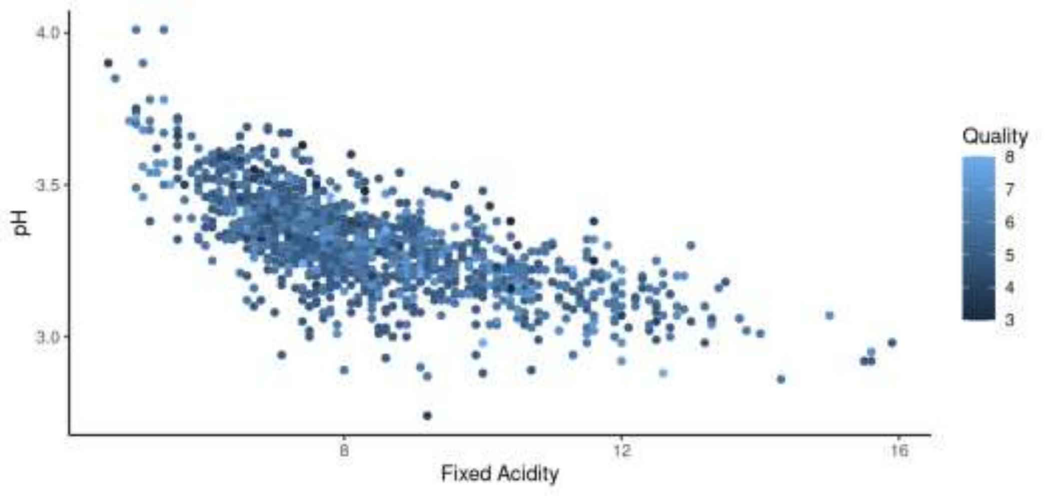
ALCOHOL CONTAINS: USING HISTOGRAM

```
install.packages(“ggplot2”)
library(ggplot2)
#for histogram of
ggplot(wine_subset,aes(x=alcohol))+geom_histogram(binwidth = 0.1,color="black",fill="blue")+labs(title="alcohol contains")
```



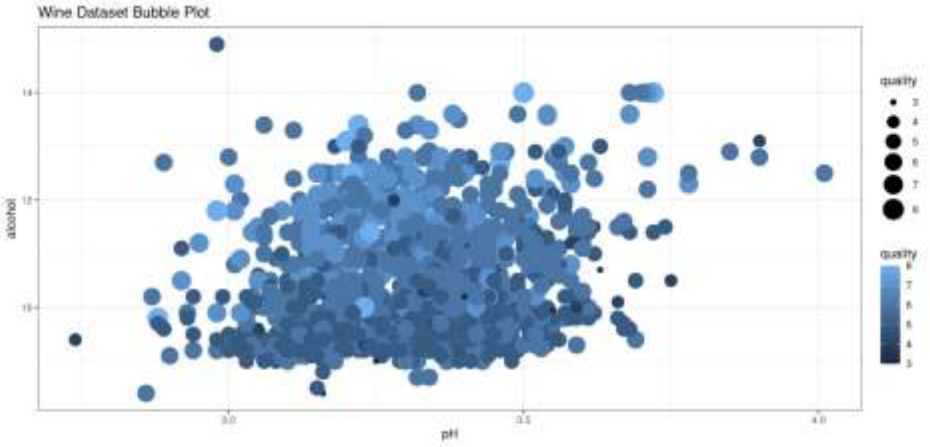
PH & FIXED ACIDITY: USING SCATTER PLOT

```
install.packages(“ggplot2”)
library(ggplot2)
# Plot fixed acidity against pH, colored by wine quality
ggplot(wine_subset, aes(x = fixed.acidity, y = pH, color = quality)) +  geom_point() +  labs(x = "Fixed Acidity", y = "pH", color =
"Quality") +  theme_classic()
```



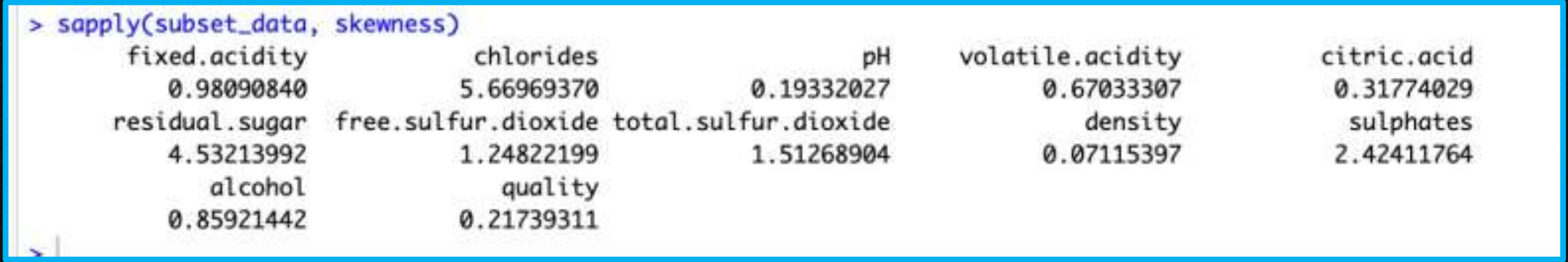
ALCOHOL QUALITY: USING BUBBLE GRAPH

```
library(ggplot2)
library(dplyr)
# Rename the columns for easier reference
colnames(wine) <- c("fixed.acidity", "chlorides", "pH", "volatile.acidity", "citric.acid","residual.sugar","free.sulfur.dioxide",
"total.sulfur.dioxide", "density","sulphates", "alcohol", "quality")
# Create the bubble plot using ggplot2
ggplot(wine, aes(x = pH, y = alcohol, size = quality, color = quality)) + geom_point() + scale_size(range = c(2, 8)) + labs(title =
"Wine Dataset Bubble Plot", x = "pH", y = "alcohol", size = "quality", color = "quality") + theme_bw()
```



skewness:

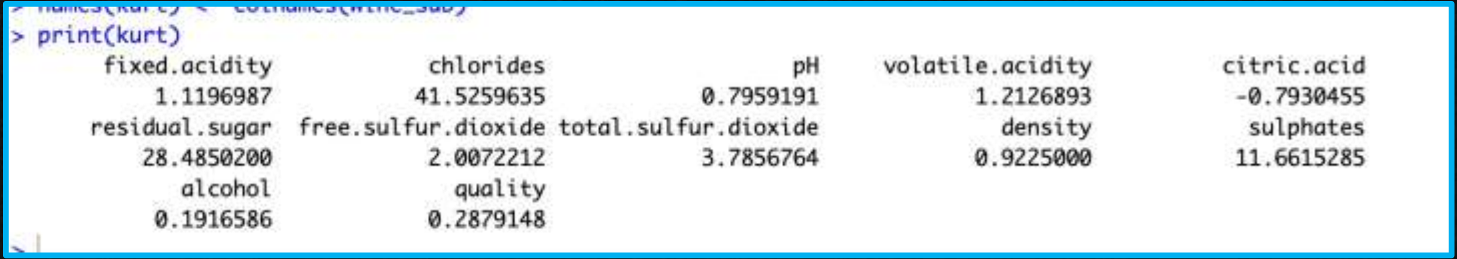
```
data(wine)
subset_data <- wine[, c("fixed.acidity", "chlorides", "pH", "volatile.acidity", "citric.acid",
"residual.sugar", "free.sulfur.dioxide","total.sulfur.dioxide", "density", "sulphates", "alcohol",
"quality")]
library(moments)
sapply(subset_data, skewness)
```



```
> sapply(subset_data, skewness)
      fixed.acidity           chlorides                  pH    volatile.acidity         citric.acid
         0.98090840          5.66969370          0.19332027          0.67033307          0.31774029
     residual.sugar free.sulfur.dioxide total.sulfur.dioxide             density           sulphates
         4.53213992          1.24822199          1.51268904          0.07115397          2.42411764
            alcohol             quality
         0.85921442          0.21739311
>
```

Kurtosis:

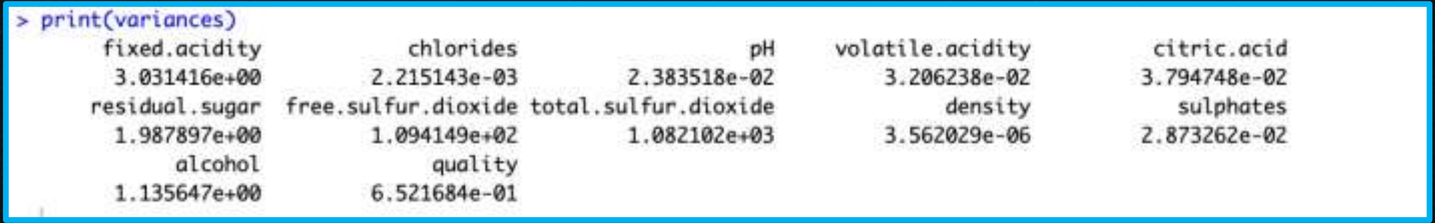
```
library(e1071)
 # for kurtosis function
# load Wine dataset
data(wine)
# select columns of interest
cols <- c( "fixed.acidity",  "chlorides",  "pH","volatile.acidity",  "citric.acid",  "residual.sugar",
"free.sulfur.dioxide",  "total.sulfur.dioxide", "density",   "sulphates",  "alcohol",  "quality")wine_sub
<- wine[, cols]
# compute kurtosis for each column
kurt <- apply(wine_sub, 2, kurtosis)
# print results
names(kurt)  <- colnames(wine_sub)
print(kurt)
```



```
> names(kurt) <- colnames(wine_sub)
> print(kurt)
       fixed.acidity            chlorides                   pH     volatile.acidity          citric.acid
           1.1196987           41.5259635            0.7959191            1.2126893           -0.7930455
      residual.sugar  free.sulfur.dioxide total.sulfur.dioxide              density            sulphates
          28.4850200            2.0072212            3.7856764            0.9225000           11.6615285
             alcohol              quality
           0.1916586            0.2879148
>
```

variance:

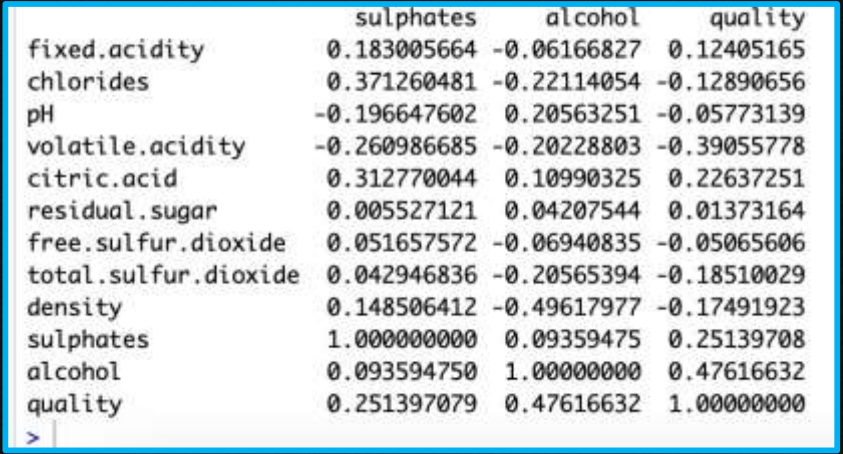
```
# load wine dataset
data(wine
)# extract the columns of interest
cols <- c("fixed.acidity", "chlorides", "pH", "volatile.acidity","citric.acid", "residual.sugar",
"free.sulfur.dioxide", "total.sulfur.dioxide", "density", "sulphates", "alcohol", "quality")wine_cols
<- wine[, cols]
# calculate the variance for each column
variances <- apply(wine_cols, 2, var)
# print the variances
print(variances)
```



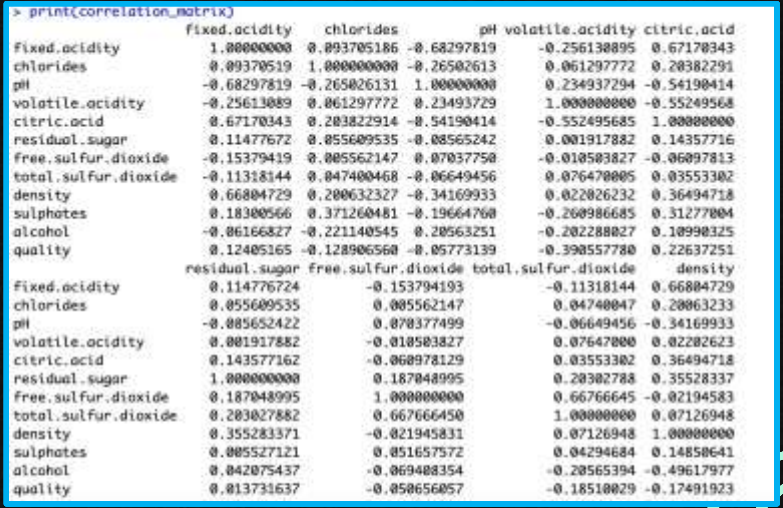
```
> print(variances)
      fixed.acidity           chlorides                  pH    volatile.acidity         citric.acid
       3.031416e+00        2.215143e-03        2.383518e-02        3.206238e-02        3.794748e-02
     residual.sugar free.sulfur.dioxide total.sulfur.dioxide           density           sulphates
       1.987897e+00        1.094149e+02        1.082102e+03        3.562029e-06        2.873262e-02
            alcohol             quality
       1.135647e+00        6.521684e-01
```

co-relation:

```
# Load the wine dataset
library(datasets)
data(wine)
# Select the columns of interest
cols <- c( "fixed.acidity",  "chlorides",  "pH","volatile.acidity",  "citric.acid",  "residual.sugar","free.sulfur.dioxide",
"total.sulfur.dioxide", "density",  "sulphates",  "alcohol",  "quality")wine_data  <- wine[, cols]
# Calculate the correlation matrix
correlation_matrix <- cor(wine_data)
# Print the correlation matrix
print(correlation_matrix)
```



```
                        sulphates     alcohol     quality
fixed.acidity         0.183005664 -0.06166827  0.12405165
chlorides             0.371260481 -0.22114054 -0.12890656
pH                   -0.196647602  0.20563251 -0.05773139
volatile.acidity     -0.260986685 -0.20228803 -0.39055778
citric.acid           0.312770044  0.10990325  0.22637251
residual.sugar        0.005527121  0.04207544  0.01373164
free.sulfur.dioxide   0.051657572 -0.06940835 -0.05065606
total.sulfur.dioxide  0.042946836 -0.20565394 -0.18510029
density               0.148506412 -0.49617977 -0.17491923
sulphates             1.000000000  0.09359475  0.25139708
alcohol               0.093594750  1.00000000  0.47616632
quality               0.251397079  0.47616632  1.00000000
>
```



```
> print(correlation_matrix)
                     fixed.acidity   chlorides          pH volatile.acidity citric.acid
fixed.acidity           1.00000000  0.093705186 -0.68297819     -0.256130895  0.67170343
chlorides               0.09370519  1.000000000 -0.26502613      0.061297772  0.20382291
pH                     -0.68297819 -0.265026131  1.00000000      0.234937294 -0.54190414
volatile.acidity       -0.25613089  0.061297772  0.23493729      1.000000000 -0.55249568
citric.acid             0.67170343  0.203822914 -0.54190414     -0.552495685  1.00000000
residual.sugar          0.11477672  0.055609535 -0.08565242      0.001917882  0.14357716
free.sulfur.dioxide    -0.15379419  0.005562147  0.07037750     -0.010503827 -0.06097813
total.sulfur.dioxide   -0.11318144  0.047400468 -0.06649456      0.076470005  0.03553302
density                 0.66804729  0.200632327 -0.34169933      0.022026232  0.36494718
sulphates               0.18300566  0.371260481 -0.19664760     -0.260986685  0.31277004
alcohol                -0.06166827 -0.221140545  0.20563251     -0.202288027  0.10990325
quality                 0.12405165 -0.128906560 -0.05773139     -0.390557780  0.22637251
                     residual.sugar free.sulfur.dioxide total.sulfur.dioxide     density
fixed.acidity           0.114776724        -0.153794193          -0.11318144  0.66804729
chlorides               0.055609535         0.005562147           0.04740047  0.20063233
pH                     -0.085652422         0.070377499          -0.06649456 -0.34169933
volatile.acidity        0.001917882        -0.010503827           0.07647000  0.02202623
citric.acid             0.143577162        -0.060978129           0.03553302  0.36494718
residual.sugar          1.000000000         0.187048995           0.20302788  0.35528337
free.sulfur.dioxide     0.187048995         1.000000000           0.66766645 -0.02194583
total.sulfur.dioxide    0.203027882         0.667666450           1.00000000  0.07126948
density                 0.355283371        -0.021945831           0.07126948  1.00000000
sulphates               0.005527121         0.051657572           0.04294684  0.14850641
alcohol                 0.042075437        -0.069408354          -0.20565394 -0.49617977
quality                 0.013731637        -0.050656057          -0.18510029 -0.17491923
```

04 conclusion

While working on Red wine Dataset in R, there are some conclusions that were seen , the dataset is all about types of physicochemical properties of contains in red wine based on the ranking of red wines(names not mentioned of grape types, wine brands or selling price)

-We saw how high total sulphur dioxide level in wine is and according to research exposure to higher concentrations can cause **nausea, vomiting, stomach pain and corrosive damage to the airways and lungs**. People with asthma may be more sensitive to the effects of sulphur dioxide.
-In alcohol histogram we can see that the mode is around 9 to 9.5 i.e from 1600 samples of wine alcohol quantity most repeated is around 9.5.
-In PH below 7 is acidic in nature & Fixed acidity is the volatility level in scatter plot the acidic is shown around 3.3 in PH and 7.2 volatility in Fixed acidity.
-In PH & alcohol bubble graph plot great quality is shown around 3.3 in PH and 9.5 in alcohol.

For conclusions, we have used various techniques and visualizations in R, such as:
Histograms, bubble plots, pie chart, scatter plot to visualize the contents of red wine in a specific amount.

THANK YOU