Kidney Stone Analysis

Patrick Udenyi

2023-9-14

In this project, I will analyze data on Kidney Stones that I found on Kaggle.com (https://www.kaggle.com/datasets/vuppalaadithyasairam/kidney-stone-prediction-based-on-urine-analysis). First I will clear my environment and load in the tidyverse package that will contain all the packages I will need for my analysis.

```r
rm(list=ls())
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

I will then read in my data I got from Kaggle by giving the path of the file on my computer relative to this file.

```r
data <- read.csv("data.csv")
```

Using the "glimpse" function I am able to see the number of rows in my data, the different columns of the data, as well as the type of data in each column.

```r
glimpse(data)
```

```
## Rows: 79
## Columns: 7
## $ gravity <dbl> 1.021, 1.017, 1.008, 1.011, 1.005, 1.020, 1.012, 1.029, 1.015,~
## $ ph      <dbl> 4.91, 5.74, 7.20, 5.51, 6.52, 5.27, 5.62, 5.67, 5.41, 6.13, 6.~
## $ osmo    <int> 725, 577, 321, 408, 187, 668, 461, 1107, 543, 779, 345, 907, 2~
## $ cond    <dbl> 14.0, 20.0, 14.9, 12.6, 7.5, 25.3, 17.4, 35.9, 21.9, 25.7, 11.~
## $ urea    <int> 443, 296, 101, 224, 91, 252, 195, 550, 170, 382, 152, 448, 64,~
## $ calc    <dbl> 2.45, 4.49, 2.36, 2.15, 1.16, 3.34, 1.40, 8.48, 1.16, 2.21, 1.~
## $ target  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

Columns correspond to the following parameters:

1.gravity - the specific gravity which is the density of urine relative to water

2.ph - pH which is the negative logarithm of the hydrogen ion concentration

3.osmo - Osmolarity (mOsm) is the proportional to the concentraion of molecules in solution

4.cond - Conductivity (mMHo milliMho) value is proportional to the concentraion of charged ions in solution

5.urea - Urea concentration in millimoles per litre

6.calc - Calcium concentration in millimoles per litre

7.target - presence of kidney stones (1 means present, 0 means absent)

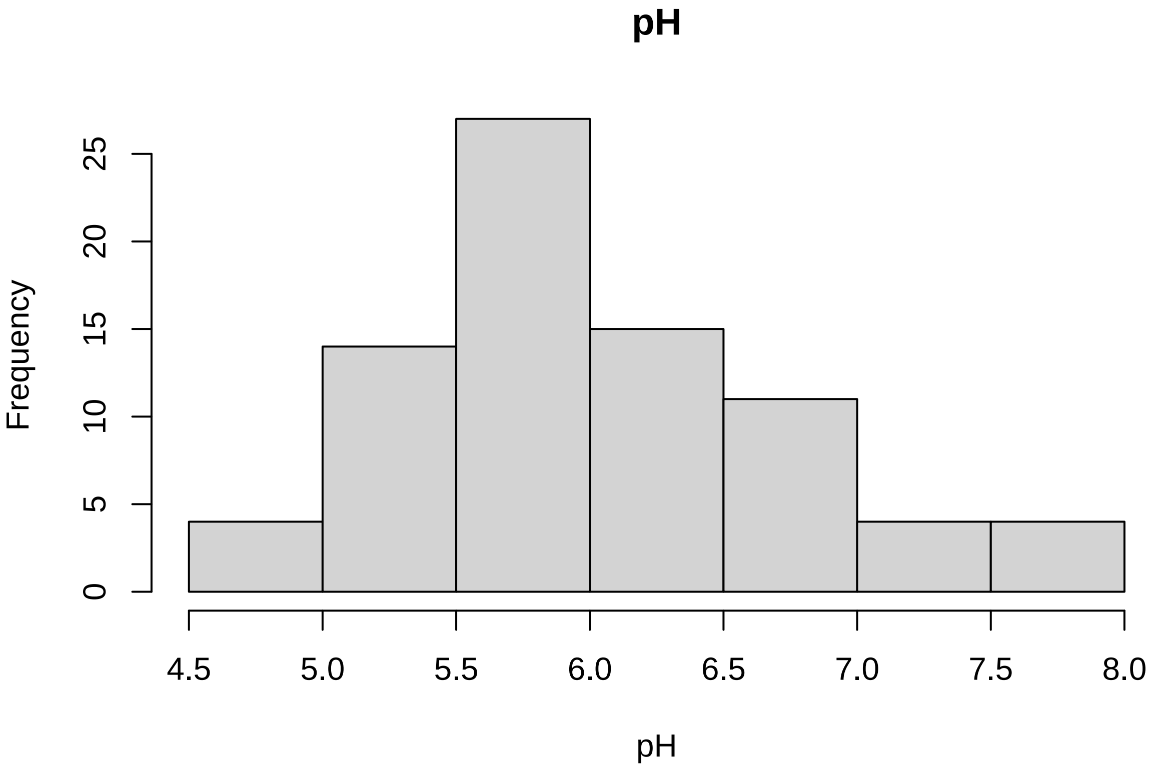
The “summary” function will allow me to get the summary statistics of each column in the dataset.

```r
summary(data)
```

```
##     gravity            ph             osmo             cond
##  Min.   :1.005   Min.   :4.760   Min.   : 187.0   Min.   : 5.10
##  1st Qu.:1.012   1st Qu.:5.530   1st Qu.: 413.0   1st Qu.:14.15
##  Median :1.018   Median :5.940   Median : 594.0   Median :21.40
##  Mean   :1.018   Mean   :6.028   Mean   : 612.8   Mean   :20.81
##  3rd Qu.:1.024   3rd Qu.:6.385   3rd Qu.: 792.0   3rd Qu.:26.55
##  Max.   :1.040   Max.   :7.940   Max.   :1236.0   Max.   :38.00
##       urea            calc            target
##  Min.   : 10.0   Min.   : 0.170   Min.   :0.0000
##  1st Qu.:160.0   1st Qu.: 1.460   1st Qu.:0.0000
##  Median :260.0   Median : 3.160   Median :0.0000
##  Mean   :266.4   Mean   : 4.139   Mean   :0.4304
##  3rd Qu.:372.0   3rd Qu.: 5.930   3rd Qu.:1.0000
##  Max.   :620.0   Max.   :14.340   Max.   :1.0000
```

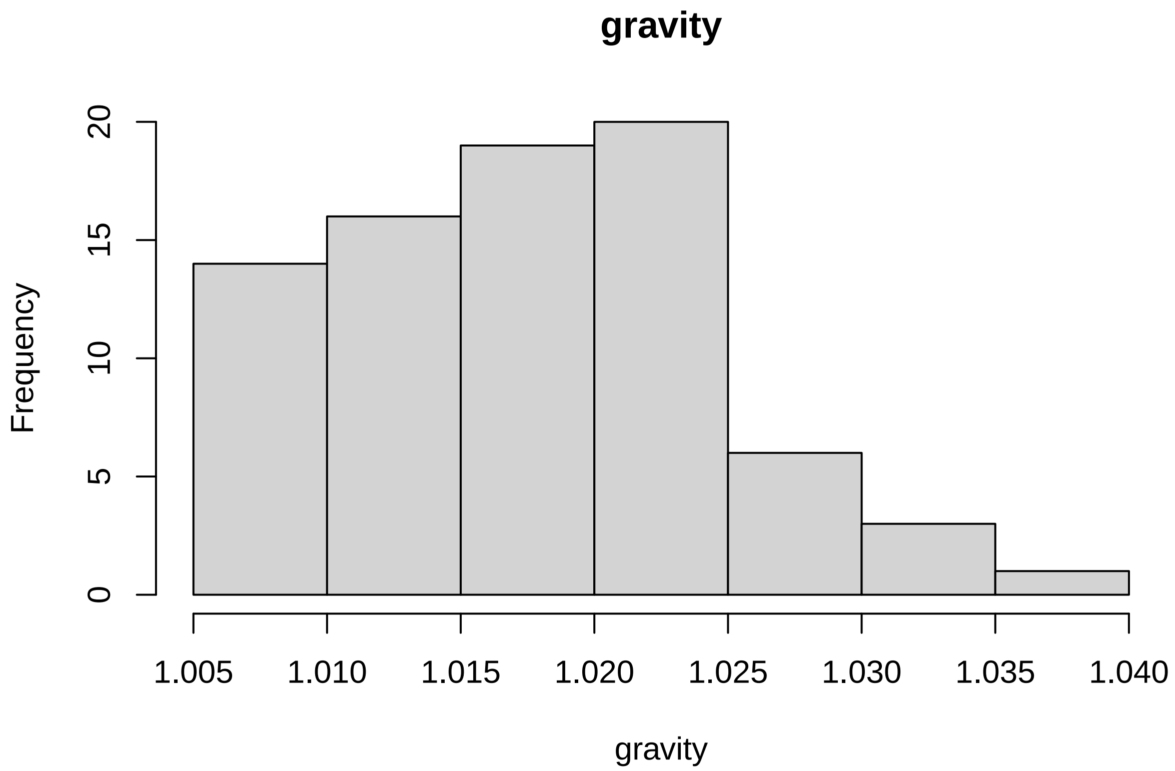
I will now create a histogram using the default “hist” function to visualize the distribution of the pH values in the dataset.

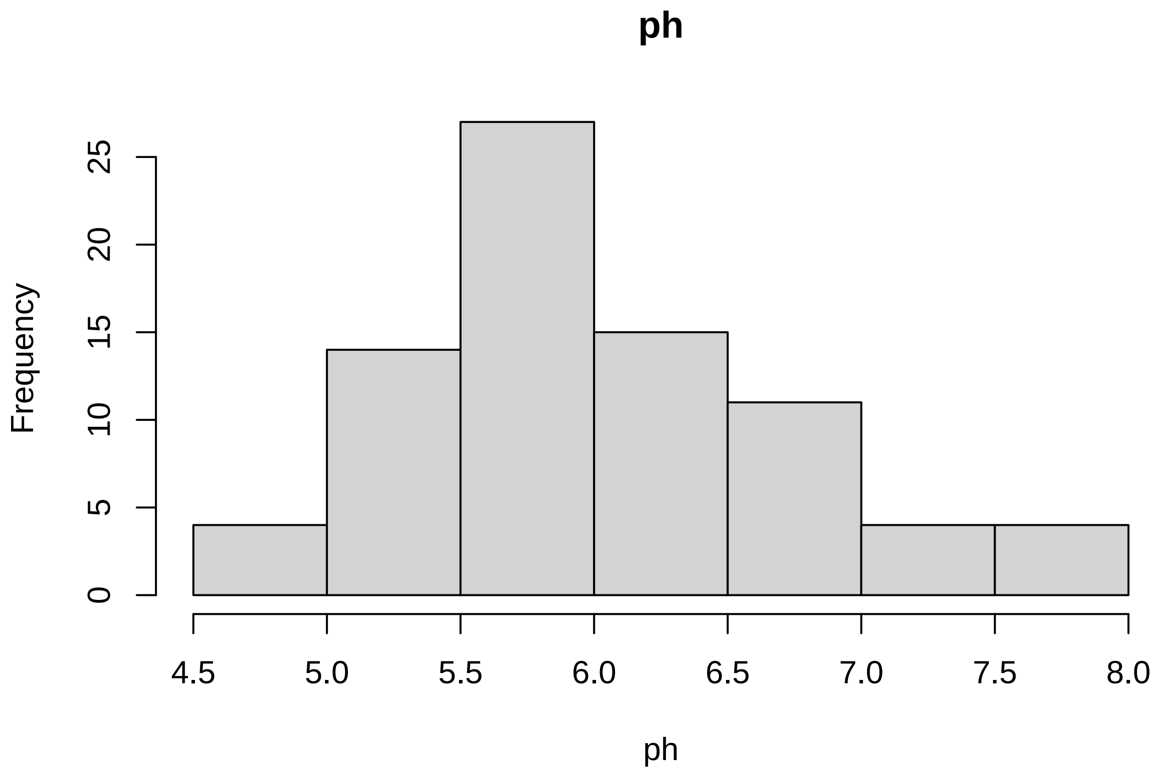
```r
hist(data$ph, main = "pH", xlab = "pH")
```

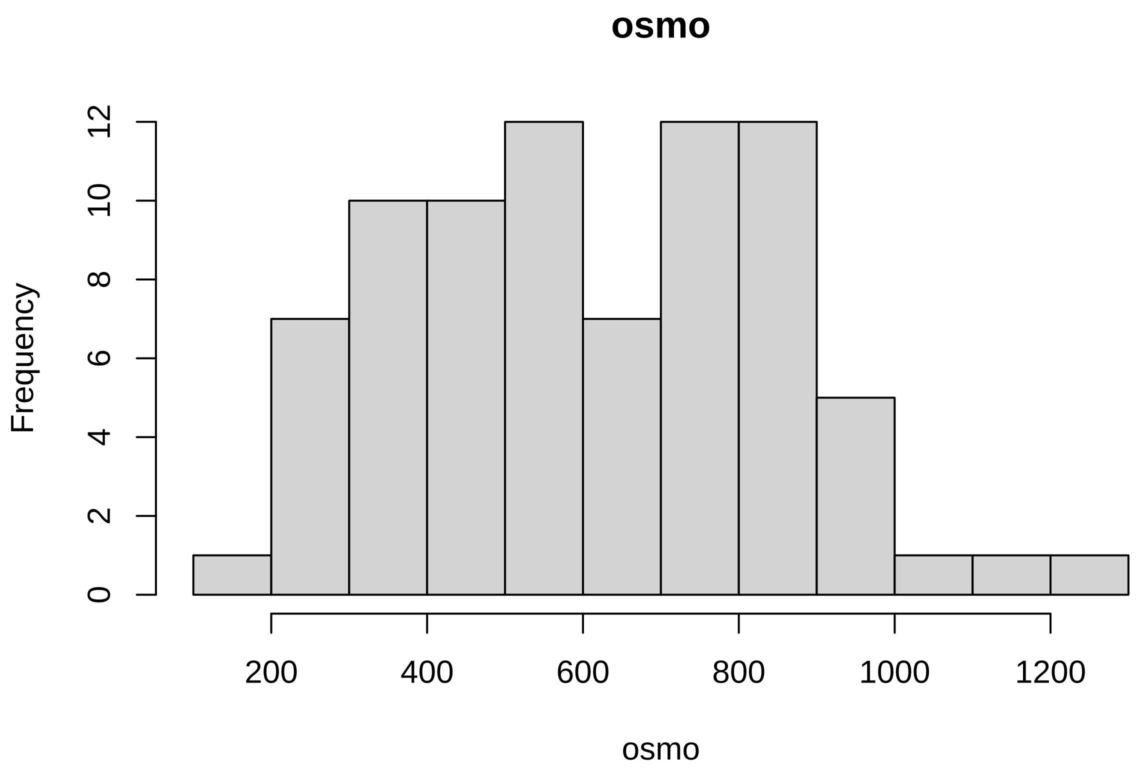


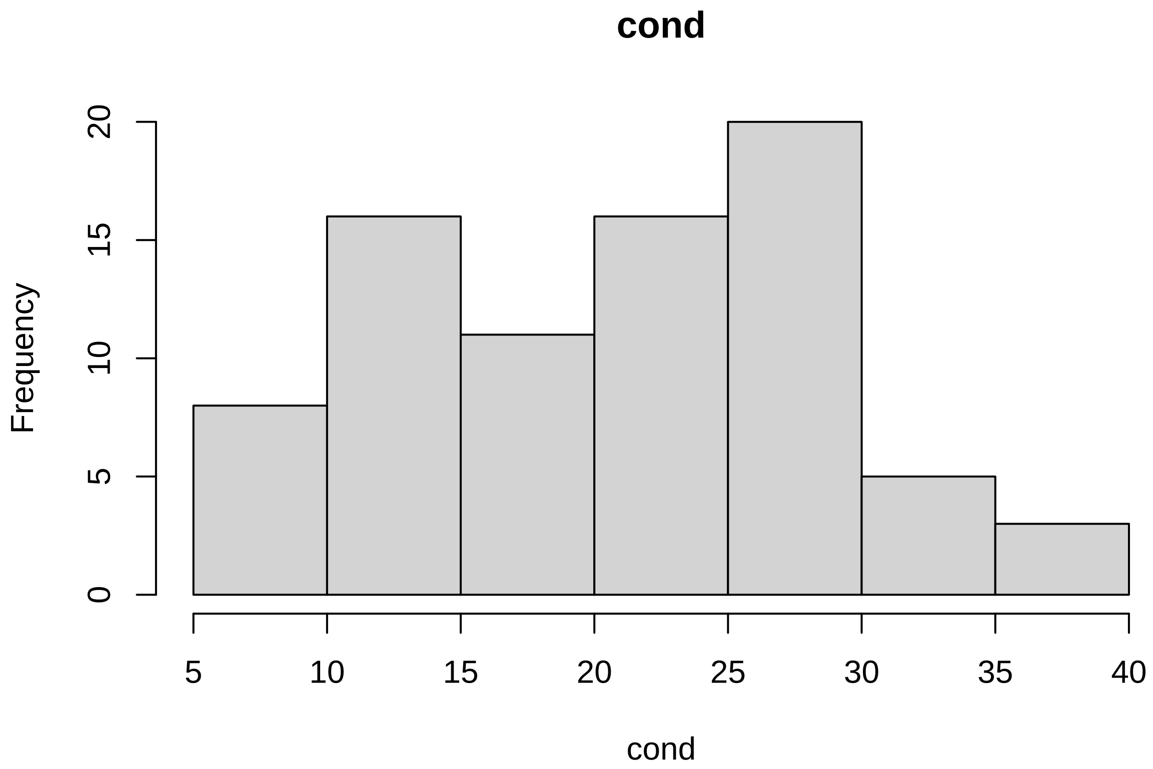
I will now generate histograms for the rest of the columns, but I don't want to repeat the same code so I will create a for loop that will iterate through each of the column names and display a histogram representing the distribution of that columns data.

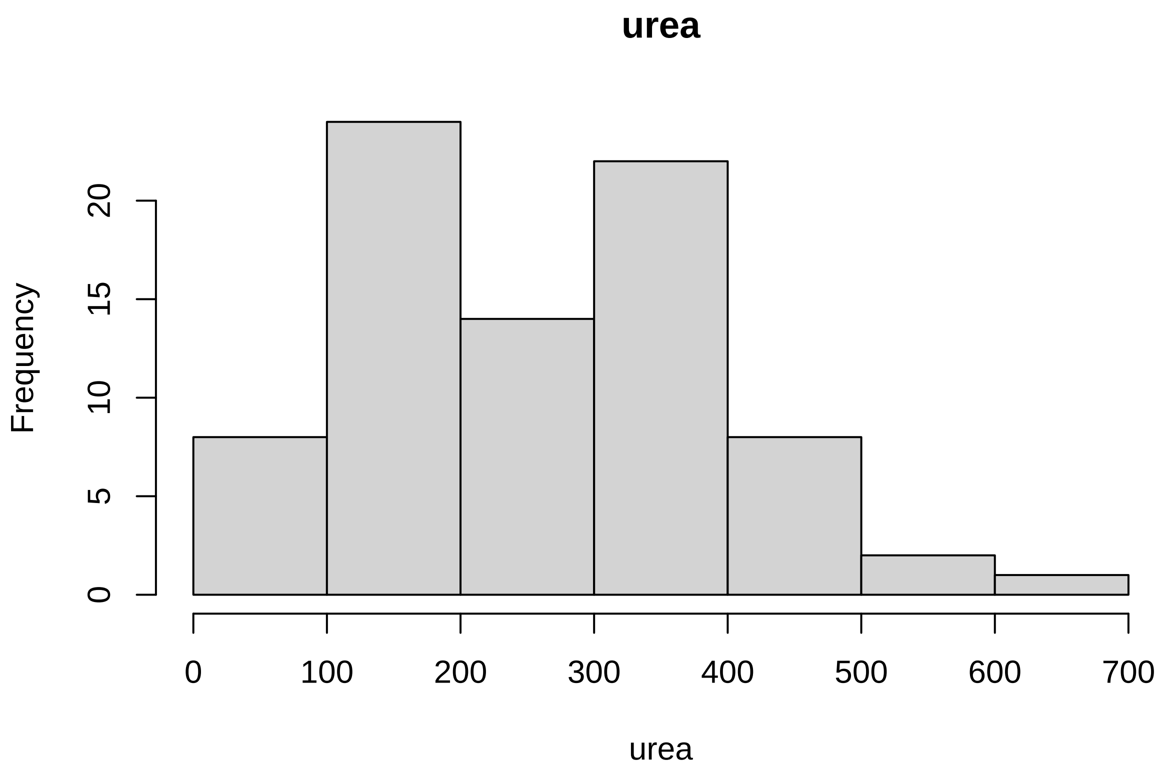
```r
for (col_name in colnames(data)) {
  hist(data[[col_name]], main = col_name, xlab = col_name)
}
```

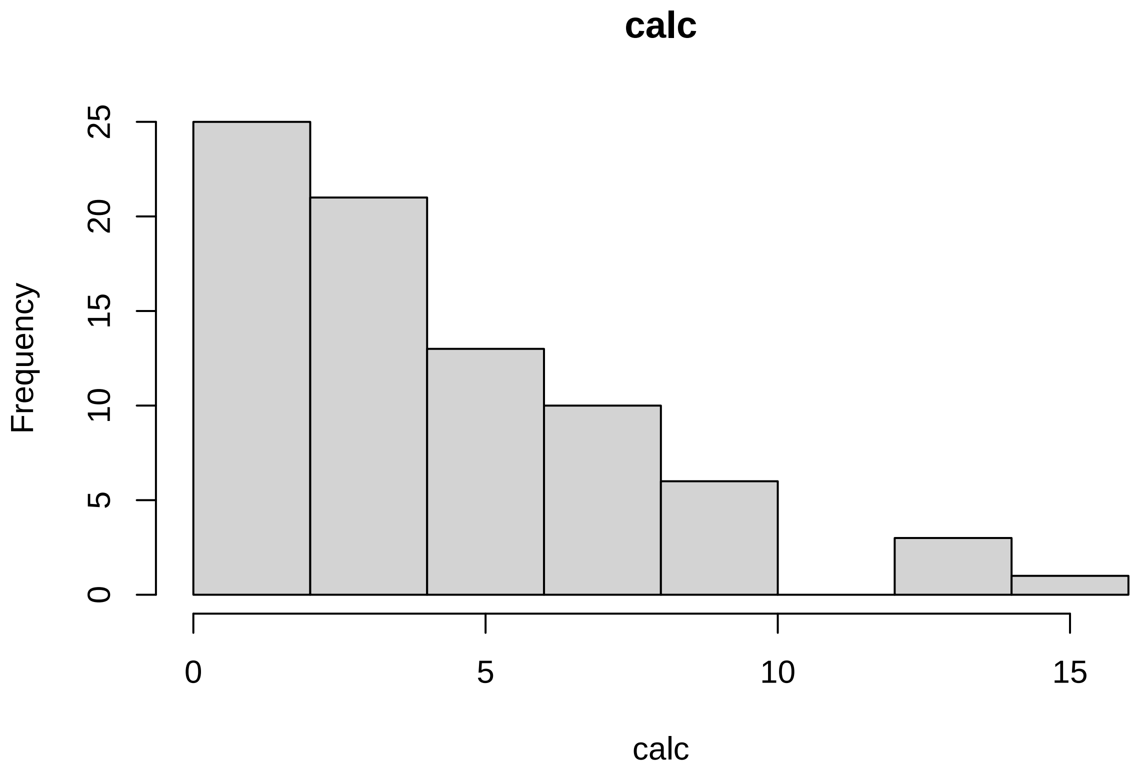

gravity
Frequency
0
5
10
15
20
1.005
1.010
1.015
1.020
1.025
1.030
1.035
1.040
gravity

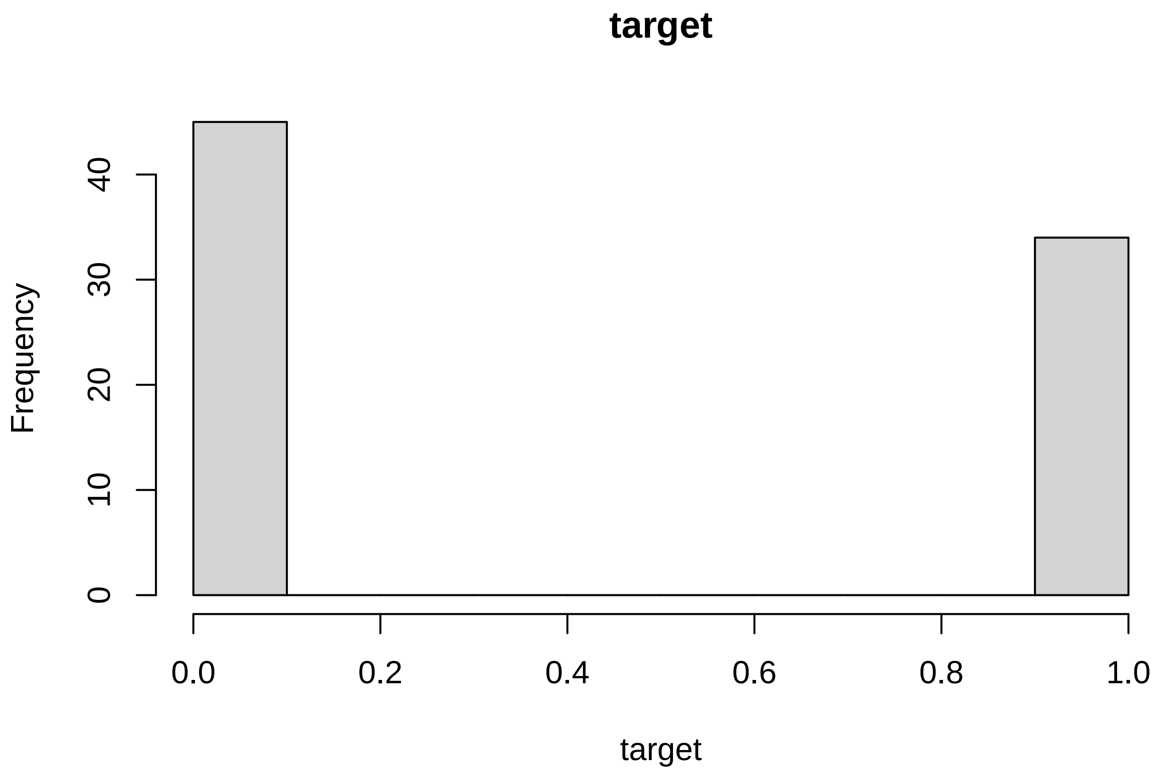

ph
Frequency
0
5
10
15
20
25
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
ph


osmo
Frequency
0
2
4
6
8
10
12
200
400
600
800
1000
1200
osmo


cond
Frequency
0
5
10
15
20
5
10
15
20
25
30
35
40
cond


urea
Frequency
0
5
10
15
20
0
100
200
300
400
500
600
700
urea


calc
Frequency
0
5
10
15
20
25
0
5
10
15
calc



In the "target" column, the values are either 1 or 0, 1 indicates a kidney stone and 0 indicates no kidney stone. I will use the "subset" function to split the data set into two different data sets to see if there is any significant differences between the pH in samples with kidney stones and sampels without kidney stones.

```
stones <- subset(data, data$target == 1)
no_stones <- subset(data, data$target == 0)
```

I will calculate the mean pH value for samples with kidney stones and the value for samples without kidney stones.

```
mean(stones$ph)
```

```
## [1] 5.935588
```

```
mean(no_stones$ph)
```

```
## [1] 6.098667
```

I will now use the "t.test" function to see if this difference is statistically significant. I will set the confidence level to 0.95.

```
t.test(stones$ph, no_stones$ph, alternative = "two.sided",
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  stones$ph and no_stones$ph
## t = -0.98097, df = 68.43, p-value = 0.3301
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4947723  0.1686155
## sample estimates:
## mean of x mean of y
##  5.935588  6.098667
```

We got a p-value of 0.3301 which means that the mean pH between samples containing kidney stones and samples not containing kidney stones are not significant because the p-value is greater than 0.05.

I am now interested in seeing if any of our parameters are significantly different between the samples with or without kidney stones. I want to avoid repeating my code so I will first create an empty vector called t_test_results.

```
t_test_results <- c()
```

I will then use a for loop to do the Welch Two Sample T-test for each column except the target column, extract the p-value, and add it to the t_test_results vector. I have printed the vector to ensure it contains our p-values.

```
for (i in 1:(length(colnames(data))-1)) {
  t_test_results[i] <- t.test(stones[,i], no_stones[,i],
                              alternative = "two.sided",
                              conf.level = 0.95)$p.value
}

t_test_results
```

```
## [1] 1.836538e-04 3.300644e-01 3.988259e-02 5.873841e-01 2.343636e-02
## [6] 5.256933e-06
```

I will now make a data frame that contains a column for the parameters in our data and the p-value the mean of that parameter between the kidney stone and no kidney stone urine sample.

```
results_df <- data.frame(
  Metric = c("gravity", "ph", "osmo", "cond", "urea", "calc"),
  P_Value = t_test_results
)

print(results_df)
```

```
##    Metric      P_Value
## 1 gravity 1.836538e-04
## 2      ph 3.300644e-01
## 3    osmo 3.988259e-02
## 4    cond 5.873841e-01
## 5    urea 2.343636e-02
## 6    calc 5.256933e-06
```

The parameters that are significantly different are specific gravity (gravity), osmolarity of the sample (osmo), urea concentration (urea), and calcium concentration (calc). The parameters that are not statistically significant are "ph" and "cond." Two types of kidney stones are calcium stones and uric acid stones, which comprise roughly 80% and 5-10% of total cases, respectively (https://www.urologyhealth.org/urology-a-z/k/kidney-stones). The significant difference in calcium concentration (calc) and urea concentration (urea) can be attributed to the chemical makeup of kidney stones. The significant difference between osmolarity (osmo) and specific gravity (gravity) may be attributed to the difference in chemical composition between urine with kidney stones and urine without kidney stones. To increase our understanding of the chemical composition of kidney stones, work to characterize and analyze differences between calcium stones and uric acid stones will be beneficial.