Assignment 1

Eric Gong

18/02/2022

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.0.5
```

```
wine_data<-read_csv("red_wine_data.csv") # Be sure this is in your current working directory
```

```
## Rows: 1599 Columns: 12


## -- Column specification -----------------------------------------------------
## Delimiter: ","
## dbl (12): fixed acidity, volatile acidity, citric acid, residual sugar, chlo...


##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
glimpse(wine_data)
```

```
## Rows: 1,599
## Columns: 12
## $ 'fixed acidity'        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7~
## $ 'volatile acidity'     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600~
## $ 'citric acid'          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00,~
## $ 'residual sugar'       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.~
## $ chlorides              <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069~
## $ 'free sulfur dioxide'  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, ~
## $ 'total sulfur dioxide' <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 10~
## $ density                <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978,~
## $ pH                     <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39,~
## $ sulphates              <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47,~
## $ alcohol                <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 1~
## $ quality                <dbl> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5,~
```

```
wine_data_chem <- wine_data %>% select(-quality)
head(wine_data_chem)
```

```
## # A tibble: 6 x 11
##   'fixed acidity' 'volatile acidity' 'citric acid' 'residual sugar' chlorides
##             <dbl>              <dbl>         <dbl>            <dbl>     <dbl>
## 1             7.4               0.7           0               1.9     0.076
## 2             7.8               0.88          0               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               0.7           0               1.9     0.076
## 6             7.4               0.66          0               1.8     0.075
## # ... with 6 more variables: free sulfur dioxide <dbl>,
## #   total sulfur dioxide <dbl>, density <dbl>, pH <dbl>, sulphates <dbl>,
## #   alcohol <dbl>
```
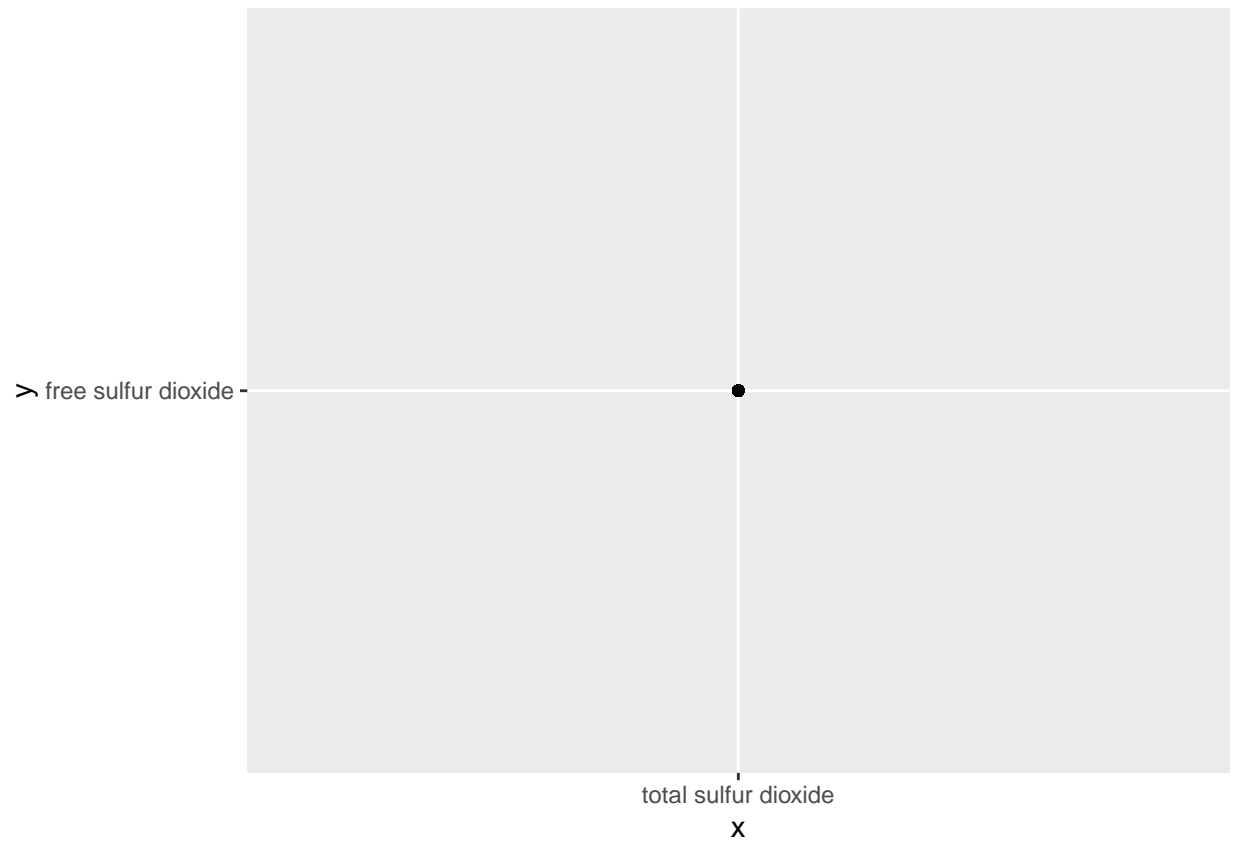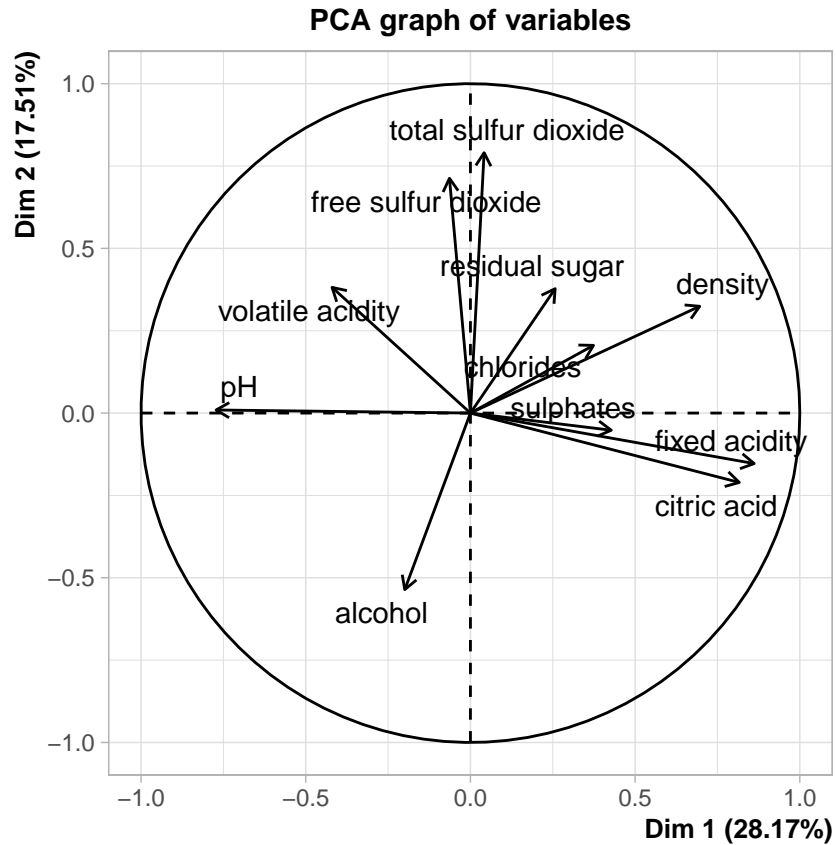
```
ggplot(wine_data_chem,aes(x="total sulfur dioxide",y="free sulfur dioxide")) +
  geom_point()
```

```
wine_pca<-PCA(wine_data_chem,graph=FALSE,ncp=11)
plot(wine_pca,choix="var")
```

PCA graph of variables


```
summary(wine_pca)
```

```
##
## Call:
## PCA(X = wine_data_chem, ncp = 11, graph = FALSE)
##
##
## Eigenvalues
##                       Dim.1   Dim.2   Dim.3    Dim.4   Dim.5   Dim.6   Dim.7
## Variance              3.099   1.926   1.551    1.213   0.959   0.660   0.584
## % of var.            28.174  17.508  14.096   11.029   8.721   5.996   5.307
## Cumulative % of var. 28.174  45.682  59.778   70.807  79.528  85.525  90.832
##                       Dim.8   Dim.9  Dim.10   Dim.11
## Variance              0.423   0.345   0.181    0.060
## % of var.             3.845   3.133   1.648    0.541
## Cumulative % of var. 94.677  97.810  99.459  100.000
##
## Individuals (the 10 first)
##                    Dist     Dim.1    ctr    cos2    Dim.2    ctr    cos2
## 1              |  2.645 |  -1.620  0.053   0.375 |  0.451  0.007  0.029 |
## 2              |  2.824 |  -0.799  0.013   0.080 |  1.857  0.112  0.432 |
## 3              |  1.936 |  -0.748  0.011   0.149 |  0.882  0.025  0.208 |
## 4              |  3.045 |   2.358  0.112   0.600 | -0.270  0.002  0.008 |
## 5              |  2.645 |  -1.620  0.053   0.375 |  0.451  0.007  0.029 |
## 6              |  2.540 |  -1.584  0.051   0.389 |  0.569  0.011  0.050 |
```

```
## 7                     |  2.115 | -1.101  0.024  0.271 |  0.608  0.012  0.083 |
## 8                     |  2.726 | -2.249  0.102  0.681 | -0.417  0.006  0.023 |
## 9                     |  2.093 | -1.087  0.024  0.270 | -0.309  0.003  0.022 |
## 10                    |  3.302 |  0.655  0.009  0.039 |  1.665  0.090  0.254 |
##                           Dim.3    ctr   cos2
## 1                        -1.774  0.127  0.450 |
## 2                        -0.912  0.034  0.104 |
## 3                        -1.171  0.055  0.366 |
## 4                         0.243  0.002  0.006 |
## 5                        -1.774  0.127  0.450 |
## 6                        -1.538  0.095  0.367 |
## 7                        -1.076  0.047  0.259 |
## 8                        -0.987  0.039  0.131 |
## 9                        -1.518  0.093  0.526 |
## 10                        1.209  0.059  0.134 |
##
## Variables (the 10 first)
##                         Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## fixed acidity        |  0.861 23.943  0.742 | -0.153  1.221  0.024 | -0.154
## volatile acidity     | -0.420  5.692  0.176 |  0.382  7.559  0.146 | -0.560
## citric acid          |  0.816 21.495  0.666 | -0.211  2.304  0.044 |  0.297
## residual sugar       |  0.257  2.135  0.066 |  0.378  7.403  0.143 |  0.126
## chlorides            |  0.374  4.505  0.140 |  0.205  2.192  0.042 | -0.115
## free sulfur dioxide  | -0.064  0.131  0.004 |  0.713 26.375  0.508 |  0.534
## total sulfur dioxide |  0.042  0.056  0.002 |  0.790 32.432  0.625 |  0.401
## density              |  0.696 15.630  0.484 |  0.324  5.456  0.105 | -0.422
## pH                   | -0.772 19.230  0.596 |  0.009  0.005  0.000 |  0.072
## sulphates            |  0.428  5.901  0.183 | -0.052  0.141  0.003 |  0.348
##                          ctr   cos2
## fixed acidity          1.520  0.024 |
## volatile acidity      20.247  0.314 |
## citric acid            5.676  0.088 |
## residual sugar         1.026  0.016 |
## chlorides              0.858  0.013 |
## free sulfur dioxide   18.386  0.285 |
## total sulfur dioxide  10.395  0.161 |
## density               11.483  0.178 |
## pH                     0.333  0.005 |
## sulphates              7.828  0.121 |
```

```r
eigenvalues_wine<-as.data.frame(wine_pca$eig) %>%
  rownames_to_column(var="Component")


eigenvalues_wine <- eigenvalues_wine %>% mutate(Component=map_chr(
  Component,~str_split(.x," ")[[1]][2]), Component_num=as.integer(Component),
  Component=factor(Component_num))


ggplot(eigenvalues_wine,aes(y=eigenvalue,x=Component_num,fill=Component)) +
geom_bar(stat="identity") + geom_line(aes(fill=NULL)) + geom_point()
```
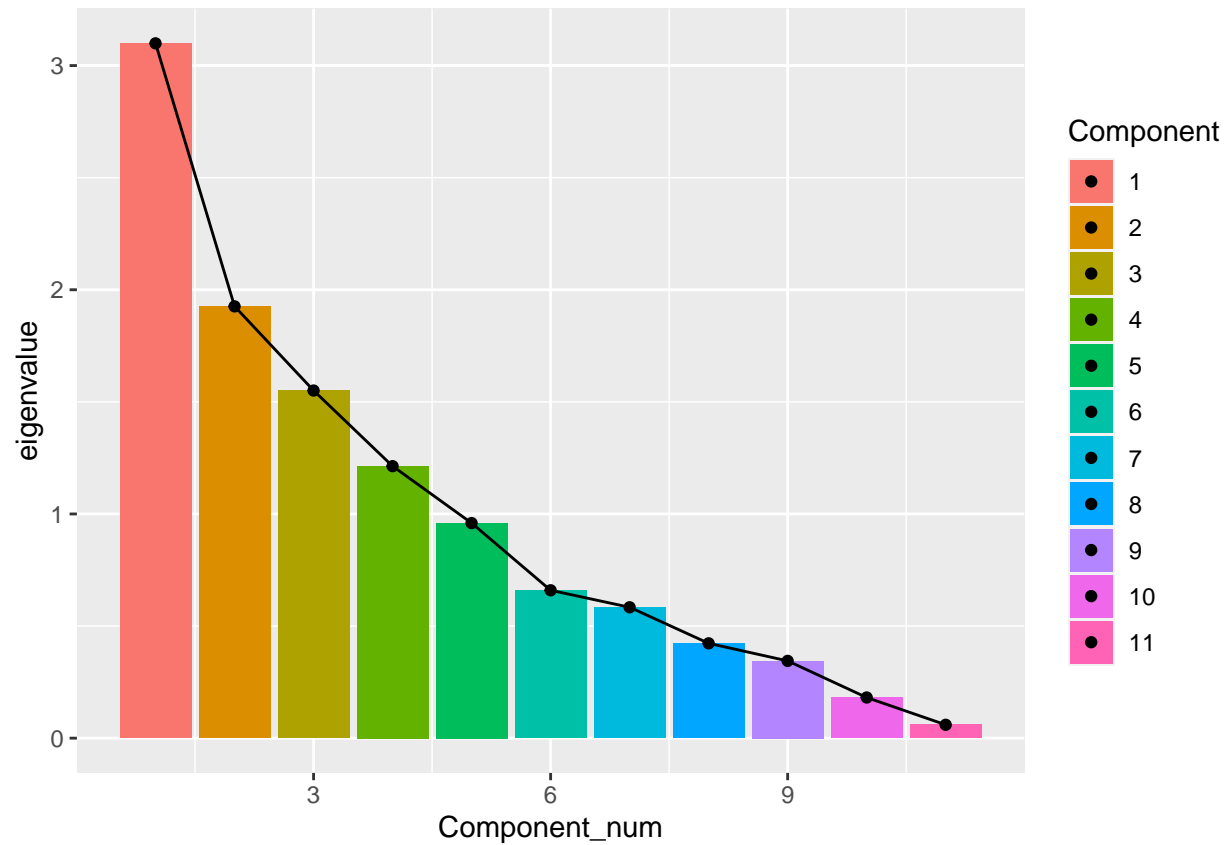
```
ggplot(eigenvalues_wine,aes(y=`cumulative percentage of variance`,
                            x=Component_num,fill=Component)) +
  geom_bar(stat="identity") + geom_line(aes(fill=NULL)) + geom_point()
```
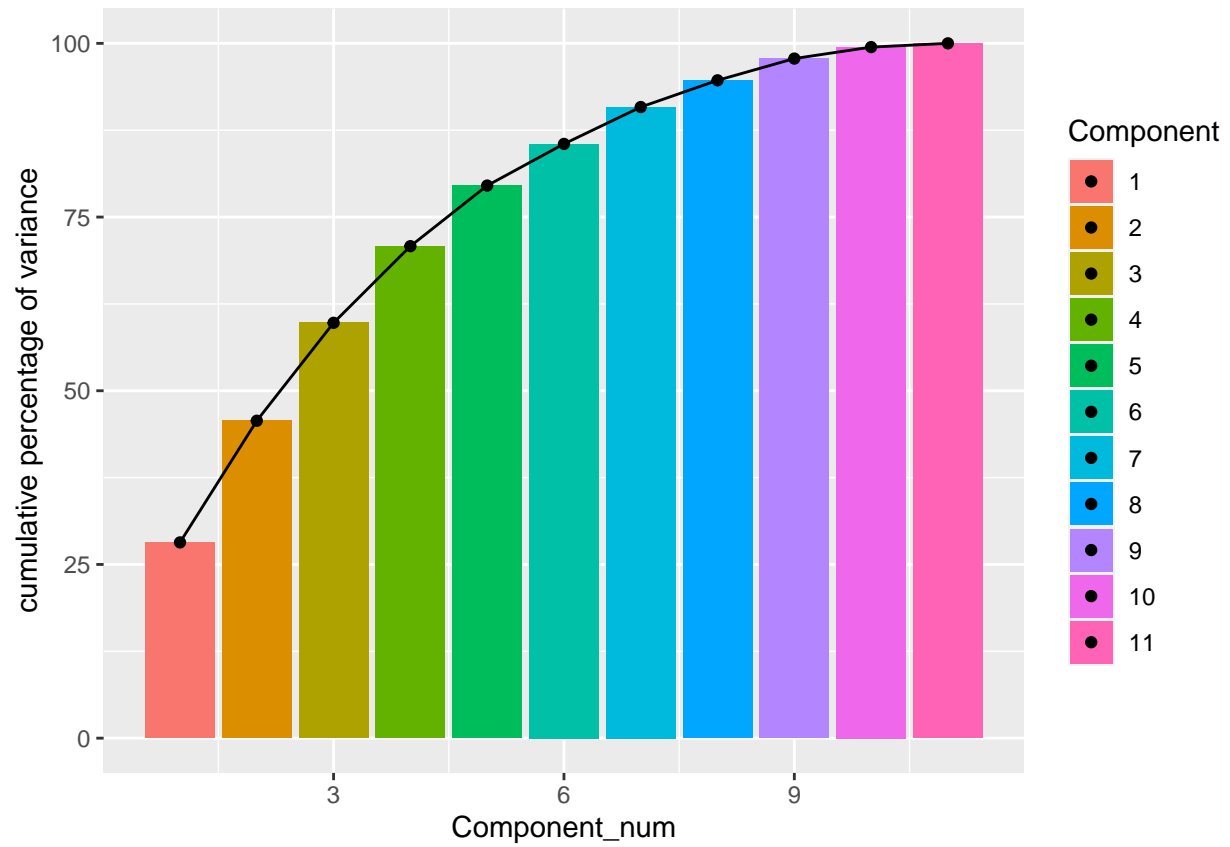
```
ggplot(eigenvalues_wine,aes(y=`percentage of variance`,x=Component_num,
                            fill=Component)) + geom_bar(stat="identity") +
  geom_line(aes(fill=NULL)) + geom_point()
```