DATAs project

DATAs

10/5/2021

First, make sure your working directory is set to the path on your desktop that contains the dataset.

You can set this by doing “Session” -> “Set Working Directory” -> “Choose Directory. . . ”

Load the csv file into our dataframe variable. (df stands for dataframe)

```
df = read_csv("StudentsPerformance.csv")
```

head() allows us to look at the first 6 rows of the data.

```
head(df)
```

```
## # A tibble: 6 x 8
##   gender ‘race/ethnicity‘ ‘parental level ~ lunch  ‘test preparati~ ‘math score‘
##   <chr>  <chr>            <chr>             <chr>  <chr>                   <dbl>
## 1 female group B          bachelor’s degree stand~ none                       72
## 2 female group C          some college      stand~ completed                  69
## 3 female group B          master’s degree   stand~ none                       90
## 4 male   group A          associate’s degr~ free/~ none                       47
## 5 male   group C          some college      stand~ none                       76
## 6 female group B          associate’s degr~ stand~ none                       71
## # ... with 2 more variables: reading score <dbl>, writing score <dbl>
```

Replace spaces between words in column names with an underscore. This will make typing the variables names out to be easier

```
colnames(df) = gsub("[^[:alnum:]]", "_", tolower(colnames(df))); colnames(df)
```

```
## [1] "gender"                      "race_ethnicity"
## [3] "parental_level_of_education" "lunch"
## [5] "test_preparation_course"     "math_score"
## [7] "reading_score"               "writing_score"
```

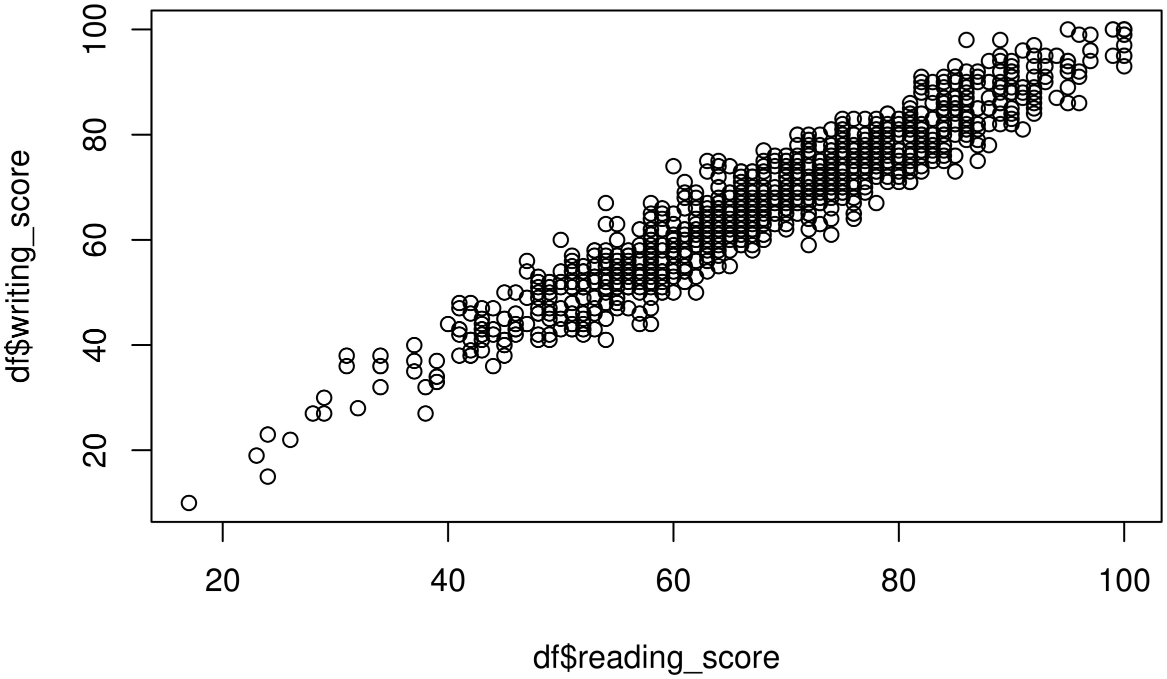
cor() gives us the correlation between the two variables we input. ‘$’ is useful for extracting the particular variable from the dataframe.

```
cor(df$reading_score, df$writing_score)
```

```
## [1] 0.9545981
```

Plotting writing score against reading score using plot(). We can see that these two variables are highly correlated!

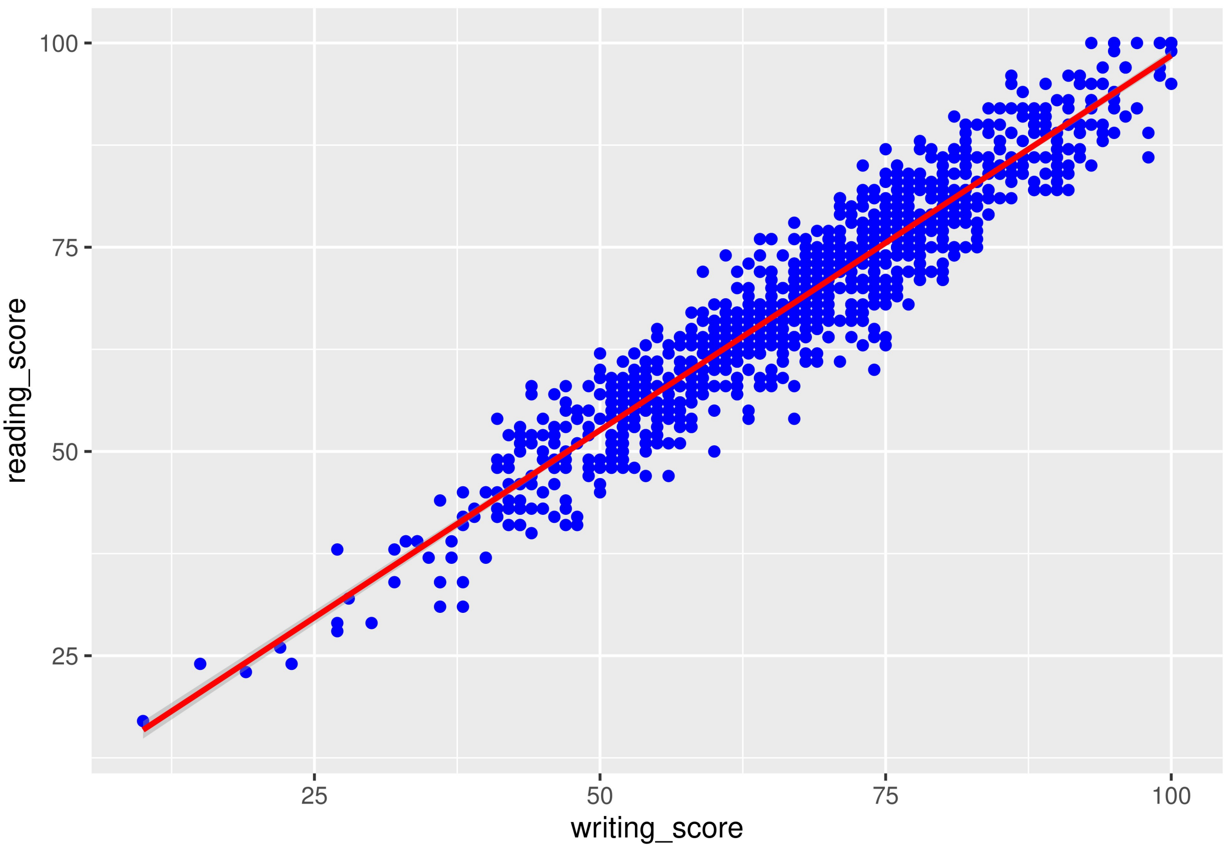
```
plot(df$reading_score, df$writing_score)
```



Plot it… but make it fancier with ggplot(). Also, add a linear regression line.

geom_point() gives the scatter plot geom_smooth() gives the linear regression line

```
ggplot(df, aes(x = writing_score, y = reading_score)) +
  geom_point(color = "blue") +
  geom_smooth(method = 'lm', formula = y ~ x, color = "red")
```


100
75
50
25
reading_score
25
50
75
100
writing_score