```
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
✓ ggplot2 3.4.1     ✓ purrr   1.0.1
✓ tibble  3.1.8     ✓ dplyr   1.1.0
✓ tidyr   1.3.0     ✓ stringr 1.5.0
✓ readr   2.1.3     ✓ forcats 1.0.0
── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> library(corrplot)
corrplot 0.92 loaded
> library(dplyr)
>
> customers <- read.csv("Customers.csv")
> view(customers)
> str(customers)
'data.frame':  2000 obs. of  10 variables:
 $ CustomerID        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender            : chr  "Male" "Male" "Female" "Female" ...
 $ Age               : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Age_numeric       : chr  "B-value" "B-value" "B-value" "B-value" ...
 $ AnnualIncome      : int  15000 35000 86000 59000 38000 58000 31000 84000 97000 980
00 ...
 $ SpendingScore     : int  39 81 6 77 40 76 6 94 3 72 ...
 $ Profession        : chr  "Healthcare" "Engineer" "Engineer" "Lawyer" ...
 $ Profession_numeric: int  1 2 2 3 4 5 1 1 2 5 ...
 $ WorkExperience    : int  1 3 1 0 2 0 1 1 0 1 ...
 $ FamilySize        : int  4 3 1 2 6 2 3 3 3 4 ...
> summary(customers)
   CustomerID        Gender               Age        Age_numeric
 Min.   :   1.0   Length:2000        Min.   : 0.00   Length:2000
 1st Qu.: 500.8   Class :character   1st Qu.:25.00   Class :character
 Median :1000.5   Mode  :character   Median :48.00   Mode  :character
 Mean   :1000.5                      Mean   :48.96
 3rd Qu.:1500.2                      3rd Qu.:73.00
 Max.   :2000.0                      Max.   :99.00
  AnnualIncome    SpendingScore     Profession        Profession_numeric
 Min.   :     0   Min.   :  0.00   Length:2000        Min.   : 1.00
 1st Qu.: 74572   1st Qu.: 28.00   Class :character   1st Qu.: 2.00
 Median :110045   Median : 50.00   Mode  :character   Median : 5.00
 Mean   :110732   Mean   : 50.96                      Mean   : 4.43
 3rd Qu.:149093   3rd Qu.: 75.00                      3rd Qu.: 5.00
 Max.   :189974   Max.   :100.00                      Max.   :10.00
 WorkExperience     FamilySize
 Min.   : 0.000   Min.   :1.000
 1st Qu.: 1.000   1st Qu.:2.000
 Median : 3.000   Median :4.000
 Mean   : 4.103   Mean   :3.768
 3rd Qu.: 7.000   3rd Qu.:5.000
 Max.   :17.000   Max.   :9.000
>
> #----------------------------Scatter Plot----------------------------------------
> ggplot(Customers,
+        aes(FamilySize,
+            SpendingScore,
+            color = Gender)) +
+   geom_point(fill ="turquoise2", size = 2)
> ggsave('P1.png')
Saving 6.52 x 5.54 in image
>
> ggplot(Customers,
+        aes(Age, WorkExperience,
```

```
+            color = Profession)) +
+   geom_point(fill= "palegreen1")
> ggsave('p2.png')
Saving 6.52 x 5.54 in image
>
> ggplot(Customers,
+        aes(SpendingScore, AnnualIncome,
+            color = Profession)) +
+   geom_point(fill= "skyblue1")
> ggsave('p3.png')
Saving 6.52 x 5.54 in image
>
> #-------------------------Box Plot-----------------------------------------
> ggplot(Customers,
+        aes(AnnualIncome,
+            Profession)) +
+   geom_boxplot(fill= "thistle2")
> ggsave('p4.png')
Saving 6.52 x 5.54 in image
>
> ggplot(Customers,
+        aes(SpendingScore,
+            Profession,
+            color = Profession)) +
+   geom_boxplot(fill ="turquoise2")
> ggsave('p5.png')
Saving 6.52 x 5.54 in image
>
> #--------------------Bar Chart-----------------------------
>
> ggplot(Customers,
+        aes(SpendingScore,
+            color = Gender)) +
+   geom_bar(fill ="plum")
> ggsave('p6.png')
Saving 6.52 x 5.54 in image
>
> #-----------------Histogram-----------------------------------
> ggplot(Customers,
+        aes(SpendingScore,
+            color = Profession)) +
+   geom_histogram(fill="snow")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggsave('p7.png')
Saving 6.52 x 5.54 in image
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
>
> # ------------------- Co-relation -----------------------------------
> customers <- read.csv("Customers.csv")
> cor.test(customers$Age, customers$FamilySize, method = "spearman")

        Spearman's rank correlation rho

data:  customers$Age and customers$FamilySize
S = 1280876395, p-value = 0.07857
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.03934246

Warning message:
In cor.test.default(customers$Age, customers$FamilySize, method = "spearman") :
  Cannot compute exact p-value with ties
> cor(select(customers, Age, FamilySize))
```

```
                    Age FamilySize
Age        1.00000000 0.03825438
FamilySize 0.03825438 1.00000000
> cus<- pairs(select(customers, Age, FamilySize,SpendingScore))
> summary(cus)
Length  Class   Mode
     0   NULL   NULL
>
> customers <- read.csv("Customers.csv")
> #customers <- filter(customers, customers$Profession_numeric == 8)
> cor.test(customers$WorkExperience, customers$FamilySize, method = "spearman")

        Spearman's rank correlation rho

data:  customers$WorkExperience and customers$FamilySize
S = 1316329885, p-value = 0.5687
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.01275234

Warning message:
In cor.test.default(customers$WorkExperience, customers$FamilySize,  :
  Cannot compute exact p-value with ties
> cor(select(customers, WorkExperience, FamilySize))
               WorkExperience FamilySize
WorkExperience     1.00000000 0.01187302
FamilySize         0.01187302 1.00000000
> pairs(select(customers, WorkExperience, FamilySize,SpendingScore))
>
> # ----------------- Linear regression --------------------------------
> customers <- read.csv("Customers.csv")
> linear_customer <- lm(customers$FamilySize ~ customers$AnnualIncome, data = custome
rs)
> summary(linear_customer)

Call:
lm(formula = customers$FamilySize ~ customers$AnnualIncome, data = customers)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0860 -1.6894 -0.0106  1.4621  5.5550

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.325e+00  1.150e-01  28.916  < 2e-16 ***
customers$AnnualIncome 4.007e-06  9.597e-07   4.175  3.1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.963 on 1998 degrees of freedom
Multiple R-squared:  0.00865,  Adjusted R-squared:  0.008154
F-statistic: 17.43 on 1 and 1998 DF,  p-value: 3.103e-05

> ggplot(customers, aes(FamilySize, AnnualIncome)) +
+   geom_point() +
+   geom_smooth(method=lm)
`geom_smooth()` using formula = 'y ~ x'
> ggsave("Linear_regression1.png")
Saving 6.52 x 5.54 in image
`geom_smooth()` using formula = 'y ~ x'
>
> #customers <- filter(customers, customers$Profession_numeric == 8 & customers$Gende
r_numeric == 1)
```

```
> linear_customer <- lm(customers$AnnualIncome ~ customers$WorkExperience, data = cus
tomers)
> summary(linear_customer)

Call:
lm(formula = customers$AnnualIncome ~ customers$WorkExperience,
    data = customers)

Residuals:
    Min      1Q  Median      3Q     Max
-108665  -36320   -1933   38272   83478

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               106467.4     1474.7    72.2  < 2e-16 ***
customers$WorkExperience   1039.5      259.9     4.0 6.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45570 on 1998 degrees of freedom
Multiple R-squared:  0.007945,  Adjusted R-squared:  0.007449
F-statistic:    16 on 1 and 1998 DF,  p-value: 6.559e-05

> ggplot(customers, aes(AnnualIncome, WorkExperience)) +
+   geom_point() +
+   geom_smooth(method=lm)
`geom_smooth()` using formula = 'y ~ x'
> ggsave("Linear_regression2.png")
Saving 6.52 x 5.54 in image
`geom_smooth()` using formula = 'y ~ x'
>
> #--------------------- T-test-------------------------------------------
> customers <- read.csv("customers.csv")
> males <- subset(customers, Gender == "Male")
> females <- subset(customers, Gender == "Female")
> t.test(males$SpendingScore, females$SpendingScore, var.equal = FALSE)

        Welch Two Sample t-test

data:  males$SpendingScore and females$SpendingScore
t = -0.023614, df = 1756.4, p-value = 0.9812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.520631  2.460656
sample estimates:
mean of x mean of y
 50.94472  50.97470

>
>
> Age_A <- subset(customers, customers$Age > 50)
> Age_B <- subset(customers, customers$Age <= 50)
> t.test(Age_A$SpendingScore, Age_B$SpendingScore, var.equal = FALSE)

        Welch Two Sample t-test

data:  Age_A$SpendingScore and Age_B$SpendingScore
t = -1.2209, df = 1990.7, p-value = 0.2223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.9743111  0.9246314
sample estimates:
mean of x mean of y
 50.16806  51.69290
```

>