EDA

Spendylove Apaloo

2024-01-24

##SALARY

```
#SALARY
library (ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.1
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
salary_data <- read.csv("salary_data_cleaned.csv")

colnames(salary_data)
```

```
##  [1] "Job.Title"         "Salary.Estimate"   "Job.Description"
##  [4] "Rating"            "Company.Name"      "Location"
##  [7] "Headquarters"      "Size"              "Founded"
## [10] "Type.of.ownership" "Industry"          "Sector"
## [13] "Revenue"           "Competitors"       "hourly"
## [16] "employer_provided" "min_salary"        "max_salary"
## [19] "avg_salary"        "company_txt"       "job_state"
## [22] "same_state"        "age"               "python_yn"
## [25] "R_yn"              "spark"             "aws"
## [28] "excel"
```

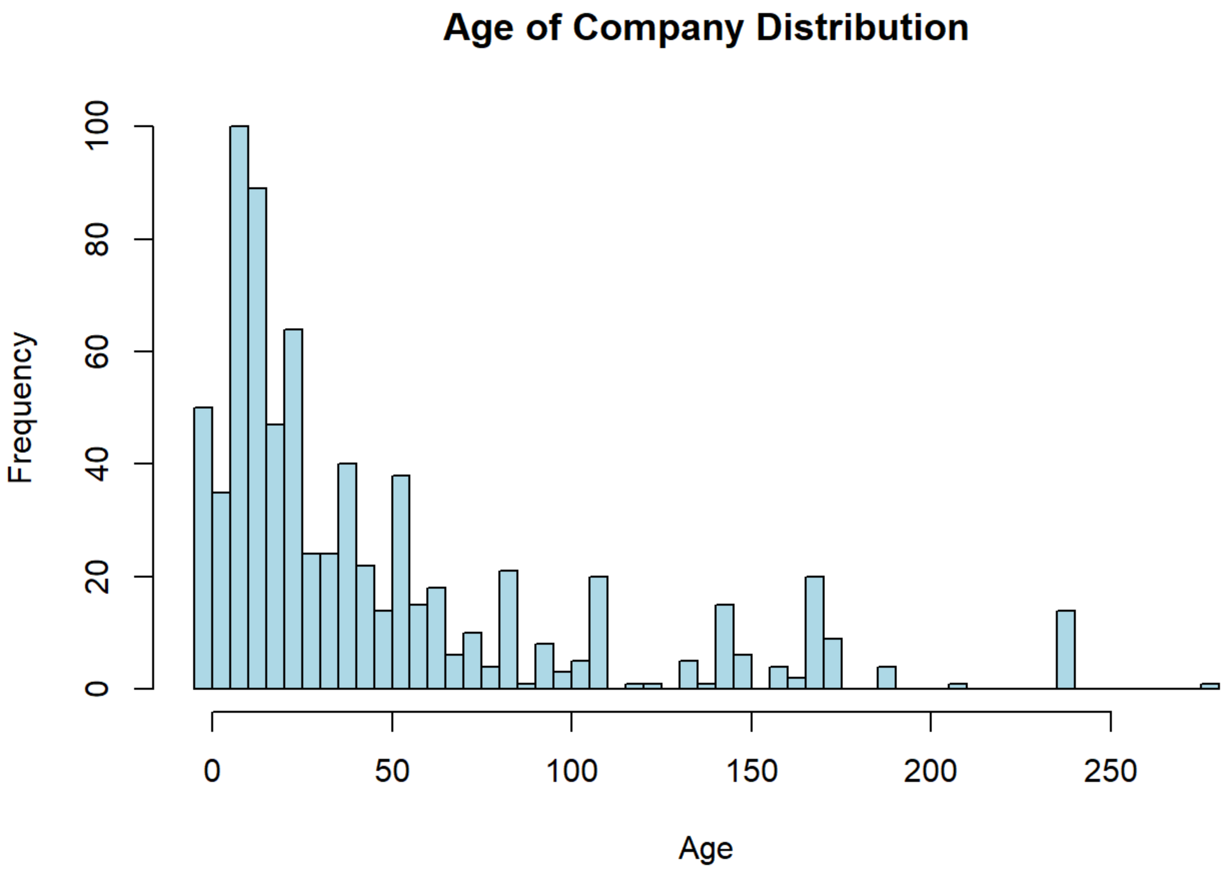
```
any(is.na(salary_data))
```

```
## [1] FALSE
```

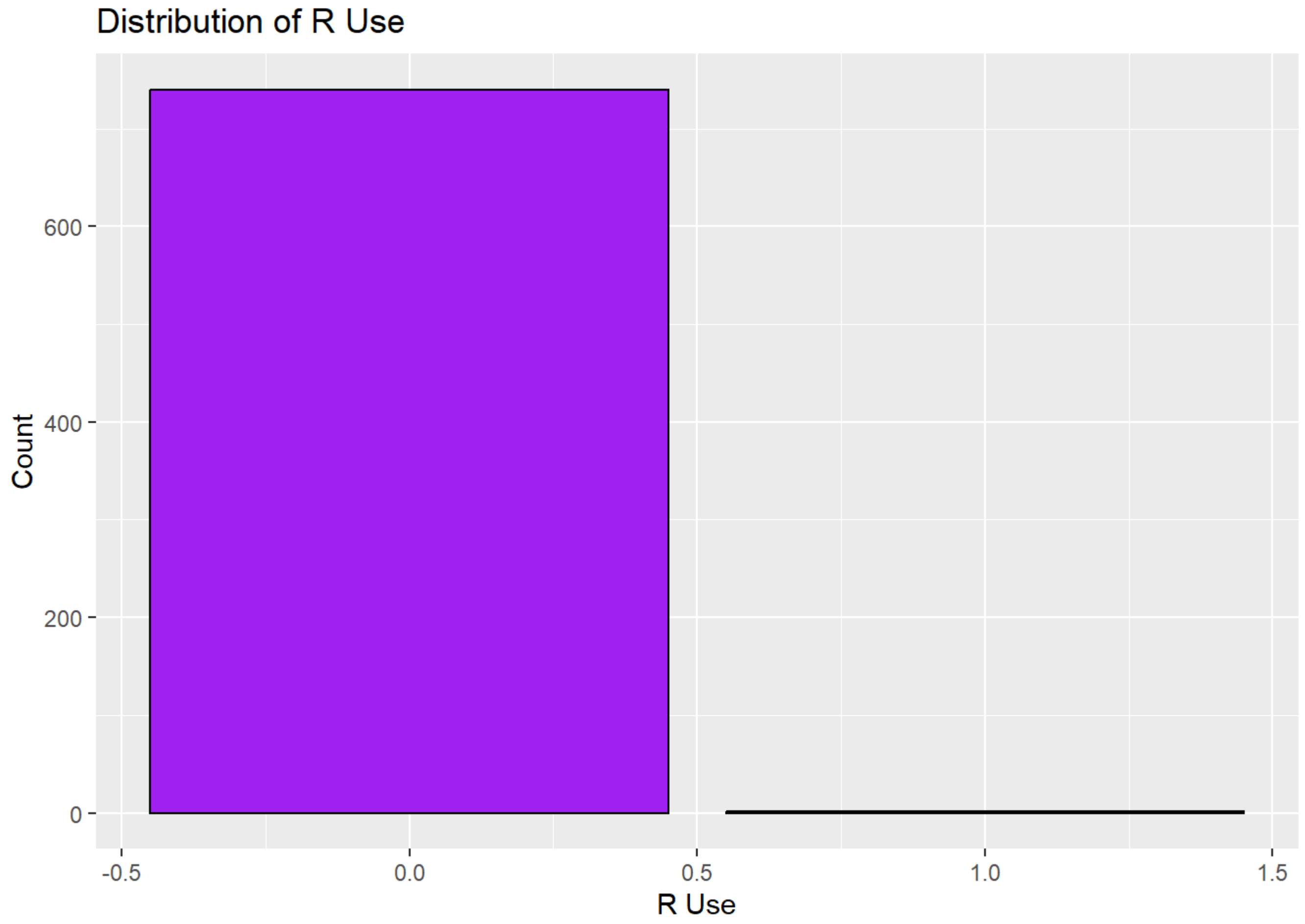
```
summary(salary_data)
```

```
##   Job.Title         Salary.Estimate    Job.Description       Rating
##  Length:742         Length:742         Length:742         Min.   :-1.000
##  Class :character   Class :character   Class :character   1st Qu.: 3.300
##  Mode  :character   Mode  :character   Mode  :character   Median : 3.700
##                                                           Mean   : 3.619
##                                                           3rd Qu.: 4.000
##                                                           Max.   : 5.000
##  Company.Name         Location         Headquarters           Size
##  Length:742         Length:742         Length:742         Length:742
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Founded     Type.of.ownership    Industry           Sector
##  Min.   :  -1   Length:742         Length:742         Length:742
##  1st Qu.:1939   Class :character   Class :character   Class :character
##  Median :1988   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1837
##  3rd Qu.:2007
##  Max.   :2019
##    Revenue           Competitors            hourly        employer_provided
##  Length:742         Length:742         Min.   :0.00000   Min.   :0.00000
##  Class :character   Class :character   1st Qu.:0.00000   1st Qu.:0.00000
##  Mode  :character   Mode  :character   Median :0.00000   Median :0.00000
##                                        Mean   :0.03234   Mean   :0.02291
##                                        3rd Qu.:0.00000   3rd Qu.:0.00000
##                                        Max.   :1.00000   Max.   :1.00000
##    min_salary       max_salary      avg_salary    company_txt
##  Min.   : 10.00   Min.   : 16.0   Min.   : 13.5   Length:742
##  1st Qu.: 52.00   1st Qu.: 96.0   1st Qu.: 73.5   Class :character
##  Median : 69.50   Median :124.0   Median : 97.5   Mode  :character
##  Mean   : 74.07   Mean   :127.2   Mean   :100.6
##  3rd Qu.: 91.00   3rd Qu.:155.0   3rd Qu.:122.5
##  Max.   :202.00   Max.   :306.0   Max.   :254.0
##   job_state           same_state         age           python_yn
##  Length:742         Min.   :0.000   Min.   : -1.00   Min.   :0.0000
##  Class :character   1st Qu.:0.000   1st Qu.: 11.00   1st Qu.:0.0000
##  Mode  :character   Median :1.000   Median : 24.00   Median :1.0000
##                     Mean   :0.558   Mean   : 46.59   Mean   :0.5283
##                     3rd Qu.:1.000   3rd Qu.: 59.00   3rd Qu.:1.0000
##                     Max.   :1.000   Max.   :276.00   Max.   :1.0000
##       R_yn              spark              aws              excel
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.002695   Mean   :0.2251   Mean   :0.2372   Mean   :0.5229
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
hist(salary_data$age, breaks = 50, col = "lightblue", border = "black",
     main = paste("Age of Company Distribution"),
     xlab = "Age", ylab = "Frequency")
```



```
ggplot(salary_data, aes(x = R_yn)) +
  geom_bar(fill = "purple", color = "black") +
  labs(title = "Distribution of R Use", x = "R Use", y = "Count")
```

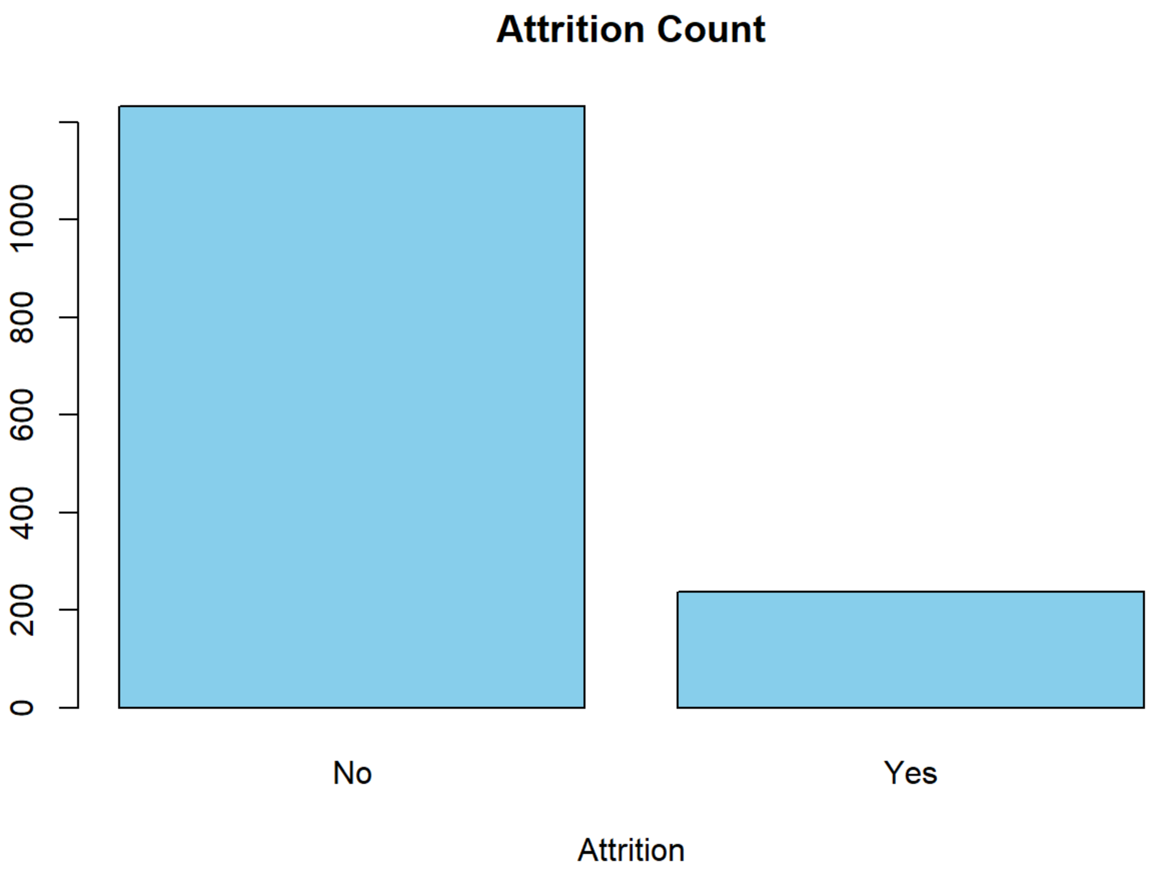


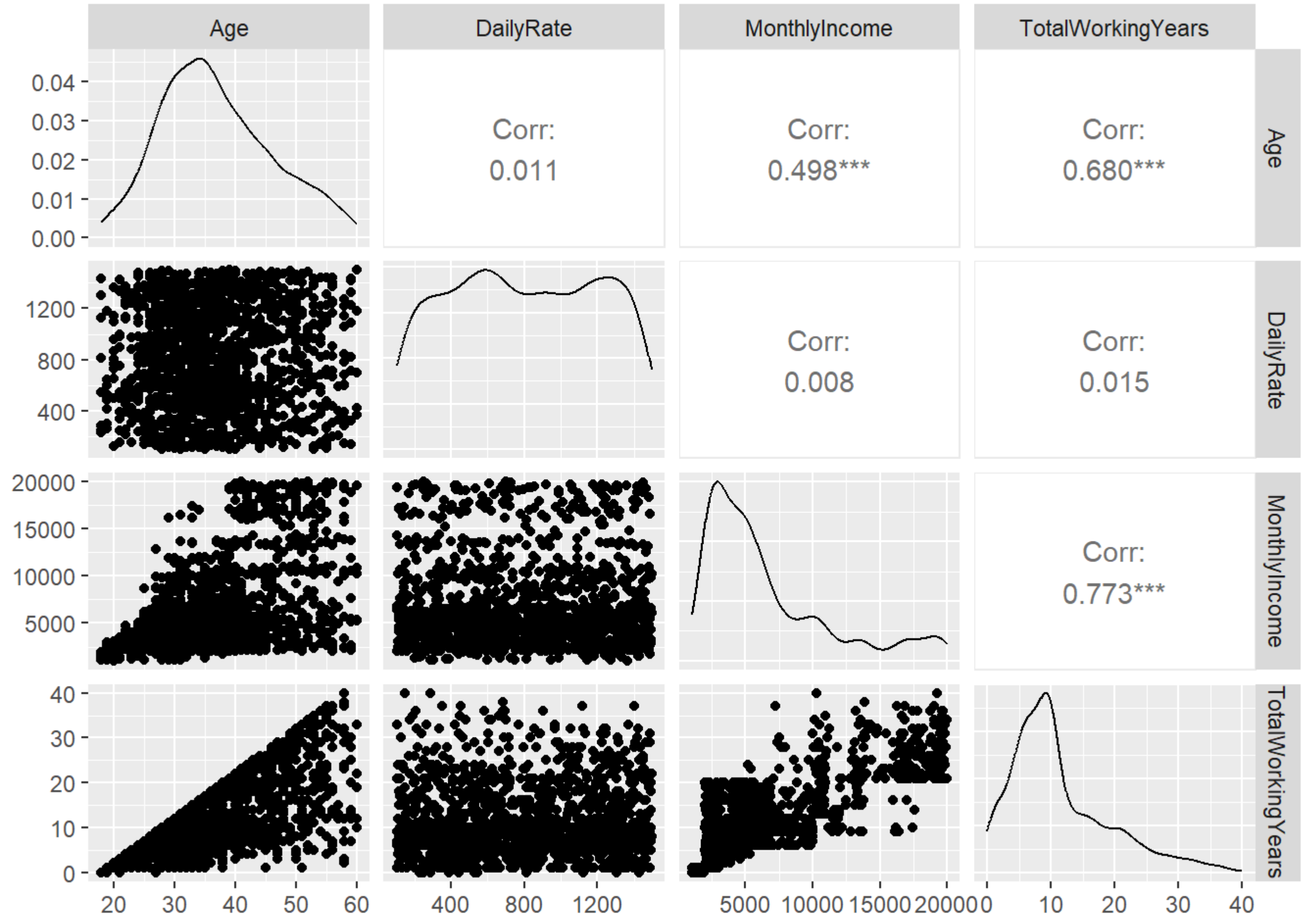
```
#Questions:
#1. Which types of companies have the highest salaries?
#2. Which states have the highest salaries and what are the job titles for these jobs?
#3. What programming language is used the most?
#4. Does company size and age have an effect of salary?
```

##HR

```
##  [1] "Age"                      "Attrition"
##  [3] "BusinessTravel"           "DailyRate"
##  [5] "Department"               "DistanceFromHome"
##  [7] "Education"                "EducationField"
##  [9] "EmployeeCount"            "EmployeeNumber"
## [11] "EnvironmentSatisfaction"  "Gender"
## [13] "HourlyRate"               "JobInvolvement"
## [15] "JobLevel"                 "JobRole"
## [17] "JobSatisfaction"          "MaritalStatus"
## [19] "MonthlyIncome"            "MonthlyRate"
## [21] "NumCompaniesWorked"       "Over18"
## [23] "OverTime"                 "PercentSalaryHike"
## [25] "PerformanceRating"        "RelationshipSatisfaction"
## [27] "StandardHours"            "StockOptionLevel"
## [29] "TotalWorkingYears"        "TrainingTimesLastYear"
## [31] "WorkLifeBalance"          "YearsAtCompany"
## [33] "YearsInCurrentRole"       "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```
## [1] FALSE
##       Age         Attrition         BusinessTravel       DailyRate
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0
##  Median :36.00   Mode  :character   Mode  :character   Median : 802.0
##  Mean   :36.92                                         Mean   : 802.5
##  3rd Qu.:43.00                                         3rd Qu.:1157.0
##  Max.   :60.00                                         Max.   :1499.0
##   Department        DistanceFromHome   Education     EducationField
##  Length:1470        Min.   : 1.000   Min.   :1.000   Length:1470
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000   Class :character
##  Mode  :character   Median : 7.000   Median :3.000   Mode  :character
##                     Mean   : 9.193   Mean   :2.913
##                     3rd Qu.:14.000   3rd Qu.:4.000
##                     Max.   :29.000   Max.   :5.000
##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender
##  Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1470
##  1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000           Class :character
##  Median :1     Median :1020.5   Median :3.000           Mode  :character
##  Mean   :1     Mean   :1024.9   Mean   :2.722
##  3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
##  Max.   :1     Max.   :2068.0   Max.   :4.000
##    HourlyRate     JobInvolvement    JobLevel       JobRole
##  Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1470
##  1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character
##  Median : 66.00   Median :3.00   Median :2.000   Mode  :character
##  Mean   : 65.89   Mean   :2.73   Mean   :2.064
##  3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
##  Max.   :100.00   Max.   :4.00   Max.   :5.000
##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate
##  Min.   :1.000   Length:1470        Min.   : 1009   Min.   : 2094
##  1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.: 8047
##  Median :3.000   Mode  :character   Median : 4919   Median :14236
##  Mean   :2.729                      Mean   : 6503   Mean   :14313
##  3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:20462
##  Max.   :4.000                      Max.   :19999   Max.   :26999
##  NumCompaniesWorked    Over18            OverTime         PercentSalaryHike
##  Min.   :0.000      Length:1470        Length:1470        Min.   :11.00
##  1st Qu.:1.000      Class :character   Class :character   1st Qu.:12.00
##  Median :2.000      Mode  :character   Mode  :character   Median :14.00
##  Mean   :2.693                                            Mean   :15.21
##  3rd Qu.:4.000                                            3rd Qu.:18.00
##  Max.   :9.000                                            Max.   :25.00
##  PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
##  Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
##  1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
##  Median :3.000     Median :3.000            Median :80    Median :1.0000
##  Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
##  3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
##  Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000
##  TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
##  Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
##  1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
##  Median :10.00     Median :3.000         Median :3.000   Median : 5.000  
##  Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008  
##  3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000  
##  Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000  
##  YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
##  Min.   : 0.000     Min.   : 0.000          Min.   : 0.000      
##  1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000      
##  Median : 3.000     Median : 1.000          Median : 3.000      
##  Mean   : 4.229     Mean   : 2.188          Mean   : 4.123      
##  3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000      
##  Max.   :18.000     Max.   :15.000          Max.   :17.000
```


Attrition Count
0
200
400
600
800
1000
No
Yes
Attrition


Age
DailyRate
MonthlyIncome
TotalWorkingYears
0.04
0.03
0.02
0.01
0.00
Corr:
0.011
Corr:
0.498***
Corr:
0.680***
Age
1200
800
400
Corr:
0.008
Corr:
0.015
DailyRate
20000
15000
10000
5000
Corr:
0.773***
MonthlyIncome
40
30
20
10
0
TotalWorkingYears
20
30
40
50
60
400
800
1200
5000
10000
15000
20000
0
10
20
30
40

```
#Questions:
#1. Is there a relationship between stock options and attrition.
#2. Does Frquency of travel have an effect on work-life balance?
#3. Do some departments experience higher rates of attrition?
#What some similarities among the departments with higher rates of attrition?
#4. Do factors such as relationship status, stock options and work like balance have an effects
of attrition rates?
```

##Shopping

```
shopping <- read.csv("shopping_trends_updated.csv")
colnames(shopping)
```

```
##  [1] "Customer.ID"            "Age"                    "Gender"
##  [4] "Item.Purchased"         "Category"               "Purchase.Amount..USD."
##  [7] "Location"               "Size"                   "Color"
## [10] "Season"                 "Review.Rating"          "Subscription.Status"
## [13] "Shipping.Type"          "Discount.Applied"       "Promo.Code.Used"
## [16] "Previous.Purchases"     "Payment.Method"         "Frequency.of.Purchases"
```

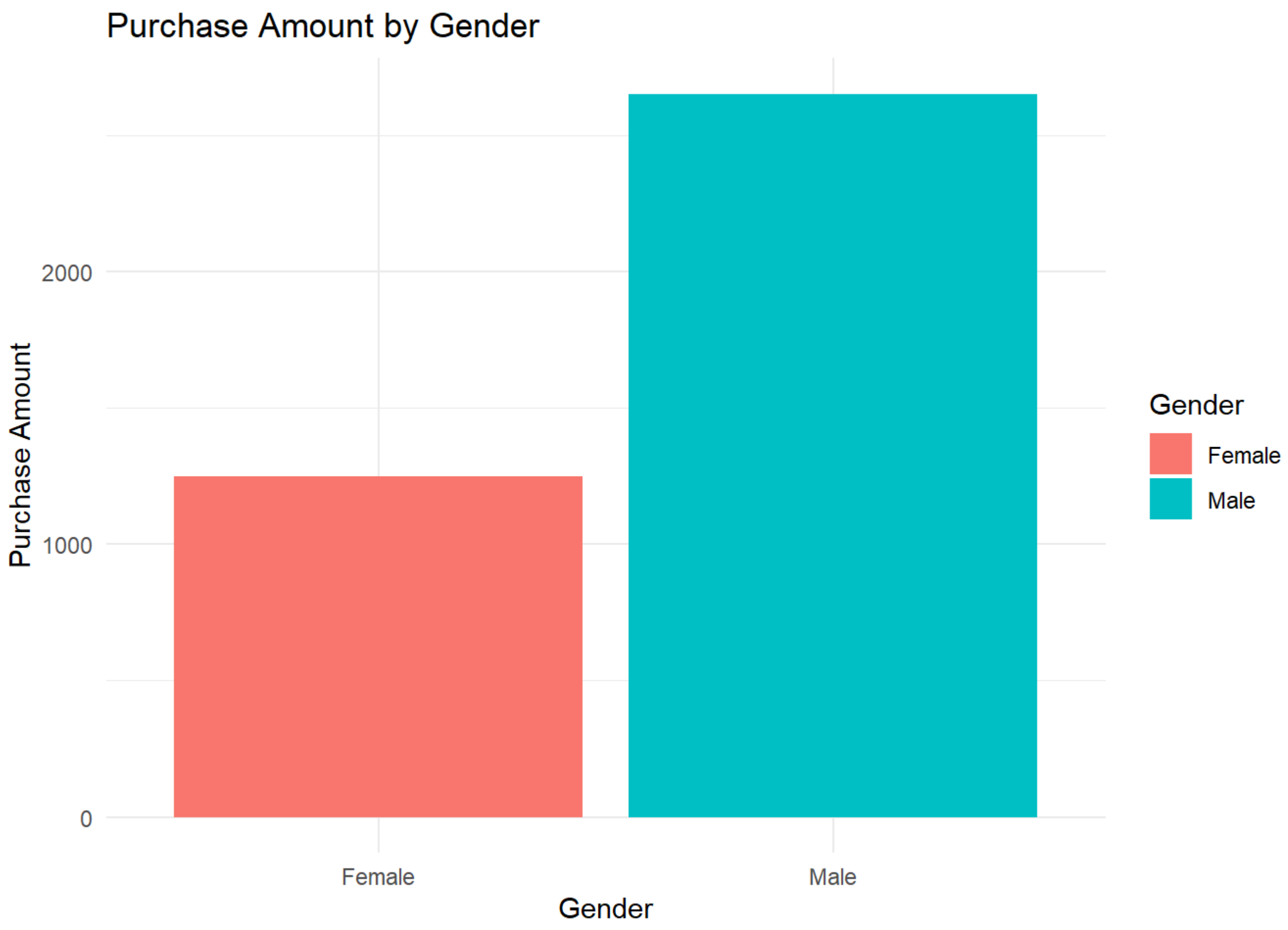
```
any(is.na(shopping))
```

```
## [1] FALSE
```

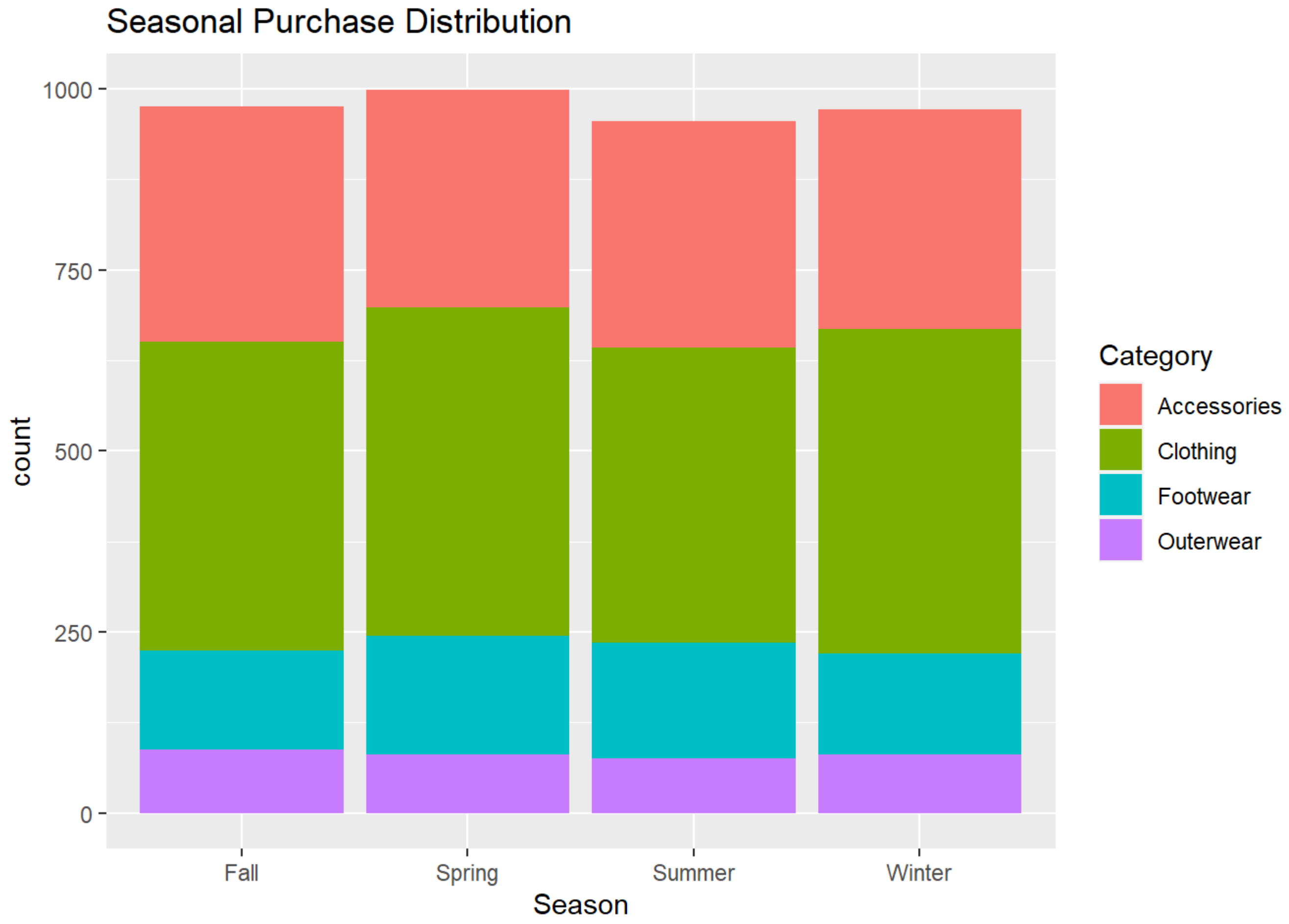
```
summary(shopping)
```

```
##   Customer.ID          Age           Gender          Item.Purchased
##  Min.   :   1.0   Min.   :18.00   Length:3900        Length:3900
##  1st Qu.: 975.8   1st Qu.:31.00   Class :character   Class :character
##  Median :1950.5   Median :44.00   Mode  :character   Mode  :character
##  Mean   :1950.5   Mean   :44.07
##  3rd Qu.:2925.2   3rd Qu.:57.00
##  Max.   :3900.0   Max.   :70.00
##    Category         Purchase.Amount..USD.   Location             Size
##  Length:3900        Min.   : 20.00        Length:3900        Length:3900
##  Class :character   1st Qu.: 39.00        Class :character   Class :character
##  Mode  :character   Median : 60.00        Mode  :character   Mode  :character
##                     Mean   : 59.76
##                     3rd Qu.: 81.00
##                     Max.   :100.00
##     Color              Season          Review.Rating  Subscription.Status
##  Length:3900        Length:3900        Min.   :2.50   Length:3900
##  Class :character   Class :character   1st Qu.:3.10   Class :character
##  Mode  :character   Mode  :character   Median :3.70   Mode  :character
##                                        Mean   :3.75
##                                        3rd Qu.:4.40
##                                        Max.   :5.00
##  Shipping.Type      Discount.Applied   Promo.Code.Used    Previous.Purchases
##  Length:3900        Length:3900        Length:3900        Min.   : 1.00
##  Class :character   Class :character   Class :character   1st Qu.:13.00
##  Mode  :character   Mode  :character   Mode  :character   Median :25.00
##                                                           Mean   :25.35
##                                                           3rd Qu.:38.00
##                                                           Max.   :50.00
##  Payment.Method     Frequency.of.Purchases
##  Length:3900        Length:3900
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
ggplot(shopping, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  labs(title = "Purchase Amount by Gender", x = "Gender", y = "Purchase Amount") +
  theme_minimal()
```



```
ggplot(shopping, aes(x = Season, fill = Category)) +
  geom_bar() +
  labs(title = "Seasonal Purchase Distribution")
```



```
#Questions:

#1. Does gender have an effect of shopping frequency?
#2. What season do people tend to shop the most?
#3. Does shipping type affect Review Rating and frequency of purchases?
#4. Which demographic shops the most, during each season?
```