Assignment1_cpendyal

2023-09-08

#Descrptive Statistics in R

#Source : https://github.com/davidcaughlin/R-Tutorial-Data-Files

#Step-1: To begin with, Set a working directory to read/write the files

#Step-2: Read the Rawfile (EmployeeSurveyDatav1.csv) and View the first 5 rows of data

```
#Read in the data
#install.packages("readr")
library(readr)

#Set Working Directory
setwd("C:\\Users\\user\\Desktop\\Masters\\Assignments\\FML Assignments\\FML Assignment 1")

surveydata <- read.csv("EmployeeSurveyDatav1.csv")
head(surveydata,5)
```

```
##   SurveyID JobSat1 JobSat2_rev JobSat3 TurnInt1 TurnInt2 TurnInt3 Engage1
## 1        1       3           3       3        3        3        3       2
## 2        2       4           2       4        3        3        2       4
## 3        3       4           2       5        2        1        2       4
## 4        4       2           3       3        4        4        4       4
## 5        5       3           3       3        4        3        3       3
##   Engage2 Engage3 Engage4 Engage5 ExpIncivil1 ExpIncivil2 ExpIncivil3
## 1       1       2       2       3           2           2           3
## 2       4       4       4       4           1           2           2
## 3       4       4       4       4           2           2           2
## 4       4       4       4       3           3           3           4
## 5       3       3       3       3           3           3           3
##   ExpIncivil4 ExpIncivil5_rev Gender Age Tenure_Yrs Location
## 1           2               4    Man  35        8.5  Seattle
## 2           3               4  Woman  42        3.9    Boise
## 3           2               4    Man  30        8.4 Portland
## 4           3               3    Man  50        1.9    Boise
## 5           3               3  Woman  56        5.2  Seattle
```

#From the above Survey data, “Gender” attributes to Categorical/Qualitative variable #From the above Survey data, “JobSat1” attributes to Quantitative variable

#Categorical/Qualitative variables are also known as Nominal Variables. Because, Nominal Variables doesn’t follow any Order.

#Quantitative variables are known as Ordinal Variables, as they have set of order in the data and an analyst can arrange the data either ascending or descending as per business requirement.

#Step-3: Measuring the Quantitative variables

```
#install.packages("dplyr") #Installed dplyr package
#install.packages("modeest")
library(modeest)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Count of Tenure_Yrs values
TenureCount <- count(surveydata, Tenure_Yrs)
TenureCount
```

```
##    Tenure_Yrs n
## 1         0.2 1
## 2         0.3 1
## 3         0.4 1
## 4         0.6 1
## 5         0.7 1
## 6         1.1 1
## 7         1.2 1
## 8         1.5 2
## 9         1.6 1
## 10        1.9 1
## 11        2.1 1
## 12        2.2 2
## 13        2.5 1
## 14        2.7 1
## 15        2.8 1
## 16        3.0 1
## 17        3.4 1
## 18        3.7 1
## 19        3.9 2
## 20        4.0 1
## 21        4.4 1
## 22        4.6 1
## 23        4.8 3
## 24        4.9 2
## 25        5.0 1
## 26        5.2 1
## 27        5.5 2
## 28        5.6 2
## 29        5.7 1
## 30        5.9 3
## 31        6.0 1
## 32        6.2 4
## 33        6.4 3
## 34        6.5 1
## 35        6.6 2
## 36        6.7 4
## 37        6.9 3
## 38        7.0 2
## 39        7.2 1
## 40        7.5 3
## 41        7.7 2
## 42        7.8 2
## 43        7.9 2
## 44        8.1 4
## 45        8.2 2
## 46        8.3 2
## 47        8.4 3
## 48        8.5 4
## 49        8.6 2
## 50        8.7 1
## 51        8.8 2
## 52        8.9 2
## 53        9.0 3
## 54        9.1 2
## 55        9.3 1
## 56        9.4 4
## 57        9.6 2
## 58        9.7 2
## 59        9.8 3
## 60       10.3 1
## 61       10.4 1
## 62       10.5 1
## 63       10.6 1
## 64       10.7 1
## 65       10.8 1
## 66       10.9 1
## 67       11.0 3
## 68       11.1 2
## 69       11.2 3
## 70       11.3 1
## 71       11.6 2
## 72       11.8 2
## 73       12.0 2
## 74       12.1 3
## 75       12.3 5
## 76       12.5 1
## 77       12.7 1
## 78       12.9 1
## 79       13.2 2
## 80       13.3 1
## 81       13.4 1
## 82       14.2 1
## 83       14.8 2
## 84       15.2 1
## 85       15.6 1
## 86       16.9 1
## 87         NA 4
```

```
#Applying Cross tab to Gender, Location
xtabs(~Gender+Location, data=surveydata)
```

```
##        Location
## Gender     Boise Portland Seattle
##          1     0        1       1
##   Man    4    15       24      42
##   Woman  4    11       17      36
```

```
#Summary of the each attribute in surveydata
summary(surveydata)
```

```
##     SurveyID         JobSat1       JobSat2_rev       JobSat3
##  Min.   :  1.00   Min.   :1.000   Min.   :1.000   Min.   :2.000
##  1st Qu.: 39.75   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:3.000
##  Median : 78.50   Median :3.000   Median :3.000   Median :3.000
##  Mean   : 78.50   Mean   :3.109   Mean   :2.763   Mean   :3.417
##  3rd Qu.:117.25   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :156.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
##                                   NA's   :4
##     TurnInt1        TurnInt2        TurnInt3        Engage1
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.250   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000
##  Median :3.000   Median :3.000   Median :3.000   Median :4.000
##  Mean   :2.981   Mean   :2.818   Mean   :2.812   Mean   :3.622
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :4.000   Max.   :5.000
##  NA's   :2       NA's   :2       NA's   :2
##     Engage2         Engage3         Engage4         Engage5
##  Min.   :1.000   Min.   :2.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :3.000   Median :4.000   Median :3.000
##  Mean   :3.378   Mean   :3.436   Mean   :3.641   Mean   :3.429
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##   ExpIncivil1     ExpIncivil2     ExpIncivil3     ExpIncivil4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.097   Mean   :2.305   Mean   :2.506   Mean   :2.565
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##  NA's   :2       NA's   :2       NA's   :2       NA's   :2
##  ExpIncivil5_rev    Gender               Age          Tenure_Yrs
##  Min.   :2.000   Length:156         Min.   :18.0   Min.   : 0.200
##  1st Qu.:3.000   Class :character   1st Qu.:33.0   1st Qu.: 5.900
##  Median :3.000   Mode  :character   Median :43.0   Median : 8.250
##  Mean   :3.513                      Mean   :43.2   Mean   : 7.998
##  3rd Qu.:4.000                      3rd Qu.:55.0   3rd Qu.:10.725
##  Max.   :5.000                      Max.   :66.0   Max.   :16.900
##  NA's   :2                          NA's   :3      NA's   :4
##    Location
##  Length:156
##  Class :character
##  Mode  :character
##
##
##
##
```

```
mfv(surveydata$JobSat1) #Most frequent value
```

```
## [1] 3
```

```
median(surveydata$JobSat1) #Median
```

```
## [1] 3
```

```
mode(surveydata$JobSat1) #Mode
```

```
## [1] "numeric"
```

```
var(surveydata$JobSat1) #Variance
```

```
## [1] 0.6654673
```

```
sd(surveydata$JobSat1) #Standard Deviation
```

```
## [1] 0.8157618
```

```
length(surveydata$JobSat1) #Length
```

```
## [1] 156
```

```
IQR(surveydata$JobSat1) #Inter Quartile Ratio
```

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

```
range(surveydata$JobSat1) #Range
```

```
## [1] 1 5
```

```
sort(surveydata$Tenure_Yrs) #Sort to Ascending
```

```
##   [1]  0.2  0.3  0.4  0.6  0.7  1.1  1.2  1.5  1.5  1.6  1.9  2.1  2.2  2.2  2.5
##  [16]  2.7  2.8  3.0  3.4  3.7  3.9  3.9  4.0  4.4  4.6  4.8  4.8  4.8  4.9  4.9
##  [31]  5.0  5.2  5.5  5.5  5.6  5.6  5.7  5.9  5.9  5.9  6.0  6.2  6.2  6.2  6.2
##  [46]  6.4  6.4  6.4  6.5  6.6  6.6  6.7  6.7  6.7  6.7  6.9  6.9  6.9  7.0  7.0
##  [61]  7.2  7.5  7.5  7.5  7.7  7.7  7.8  7.8  7.9  7.9  8.1  8.1  8.1  8.1  8.2
##  [76]  8.2  8.3  8.3  8.4  8.4  8.4  8.5  8.5  8.5  8.5  8.6  8.6  8.7  8.8  8.8
##  [91]  8.9  8.9  9.0  9.0  9.0  9.1  9.1  9.3  9.4  9.4  9.4  9.4  9.6  9.6  9.7
## [106]  9.7  9.8  9.8  9.8 10.3 10.4 10.5 10.6 10.7 10.8 10.9 11.0 11.0 11.0 11.1
## [121] 11.1 11.2 11.2 11.2 11.3 11.6 11.6 11.8 11.8 12.0 12.0 12.1 12.1 12.1 12.3
## [136] 12.3 12.3 12.3 12.3 12.5 12.7 12.9 13.2 13.2 13.3 13.4 14.2 14.8 14.8 15.2
## [151] 15.6 16.9
```

```
sort(surveydata$Tenure_Yrs, decreasing = TRUE) #Sort to Descending
```

```
##   [1] 16.9 15.6 15.2 14.8 14.8 14.2 13.4 13.3 13.2 13.2 12.9 12.7 12.5 12.3 12.3
##  [16] 12.3 12.3 12.3 12.1 12.1 12.1 12.0 12.0 11.8 11.8 11.6 11.6 11.3 11.2 11.2
##  [31] 11.2 11.1 11.1 11.0 11.0 11.0 10.9 10.8 10.7 10.6 10.5 10.4 10.3  9.8  9.8
##  [46]  9.8  9.7  9.7  9.6  9.6  9.4  9.4  9.4  9.4  9.3  9.1  9.1  9.0  9.0  9.0
##  [61]  8.9  8.9  8.8  8.8  8.7  8.6  8.6  8.5  8.5  8.5  8.5  8.4  8.4  8.4  8.3
##  [76]  8.3  8.2  8.2  8.1  8.1  8.1  8.1  7.9  7.9  7.8  7.8  7.7  7.7  7.5  7.5
##  [91]  7.5  7.2  7.0  7.0  6.9  6.9  6.9  6.7  6.7  6.7  6.7  6.6  6.6  6.5  6.4
## [106]  6.4  6.4  6.2  6.2  6.2  6.2  6.0  5.9  5.9  5.9  5.7  5.6  5.6  5.5  5.5
## [121]  5.2  5.0  4.9  4.9  4.8  4.8  4.8  4.6  4.4  4.0  3.9  3.9  3.7  3.4  3.0
## [136]  2.8  2.7  2.5  2.2  2.2  2.1  1.9  1.6  1.5  1.5  1.2  1.1  0.7  0.6  0.4
## [151]  0.3  0.2
```

#Step-4: Measuring the Nominal/Categorical variables.

```
library("dplyr")
GenderCount <- count(surveydata, Gender)
(GenderCount)
```

```
##   Gender  n
## 1         3
## 2    Man 85
## 3  Woman 68
```

```
#Describe the proportions of the Location
prop.table(table(surveydata$Location))*100
```

```
##
##               Boise  Portland   Seattle
##  5.769231 16.666667 26.923077 50.641026
```

#Step-5: Transforming a Quantitative variable

```
#Transformation
#install.packages("ggplot2") #Installed ggplot2 package
library(ggplot2)
#install.packages("gridExtra")
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
summary(surveydata$JobSat1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.109   4.000   5.000
```

```
summary(log10(surveydata$JobSat1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.4771  0.4771  0.4756  0.6021  0.6990
```

```
summary(sqrt(surveydata$JobSat1))
```

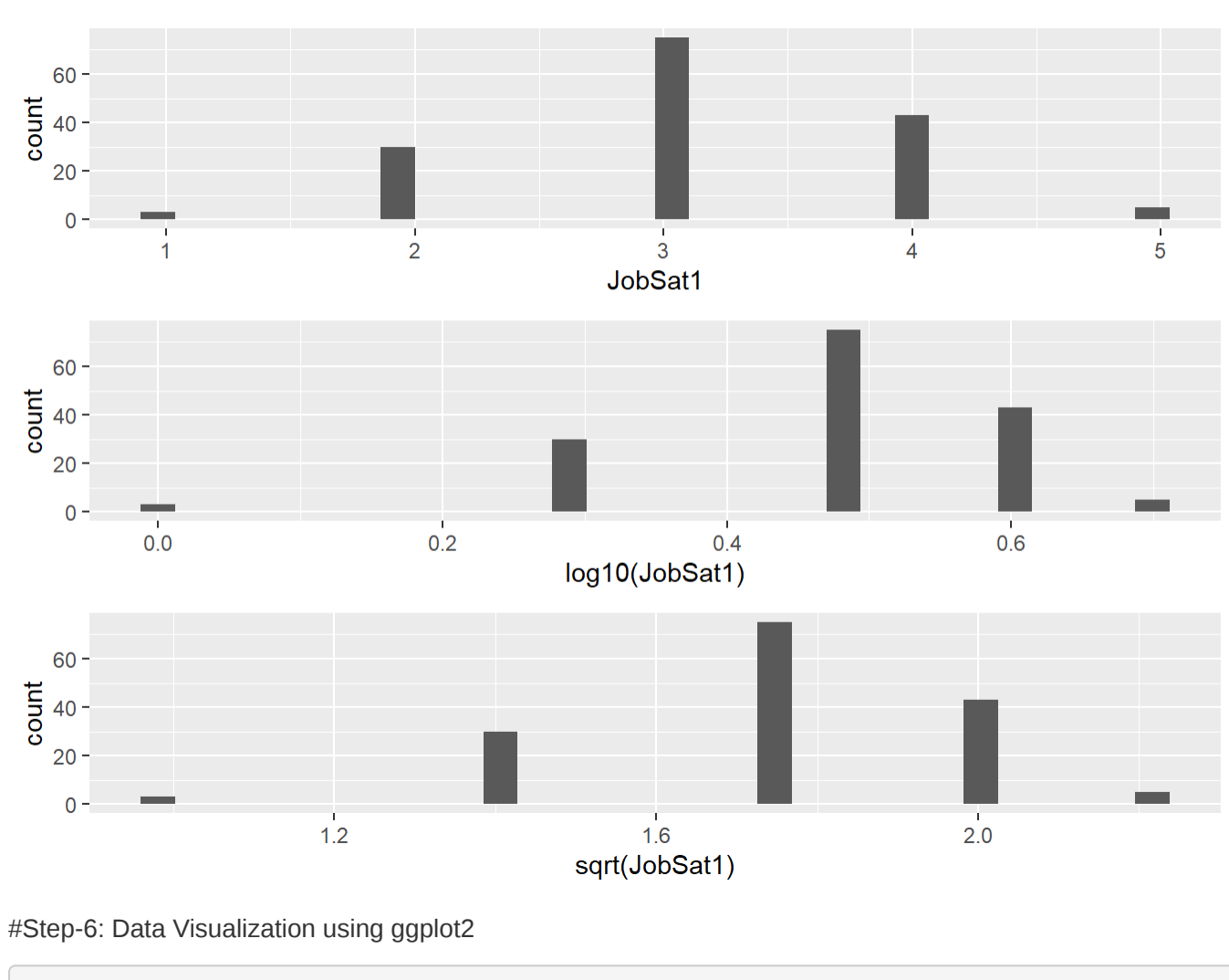
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.732   1.732   1.747   2.000   2.236
```

```
p1 <- ggplot(aes(x=JobSat1), data=surveydata) + geom_histogram()
p2 <- ggplot(aes(x=log10(JobSat1)), data=surveydata) + geom_histogram()
p3 <- ggplot(aes(x=sqrt(JobSat1)), data=surveydata) + geom_histogram()

grid.arrange(p1, p2, p3, ncol=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

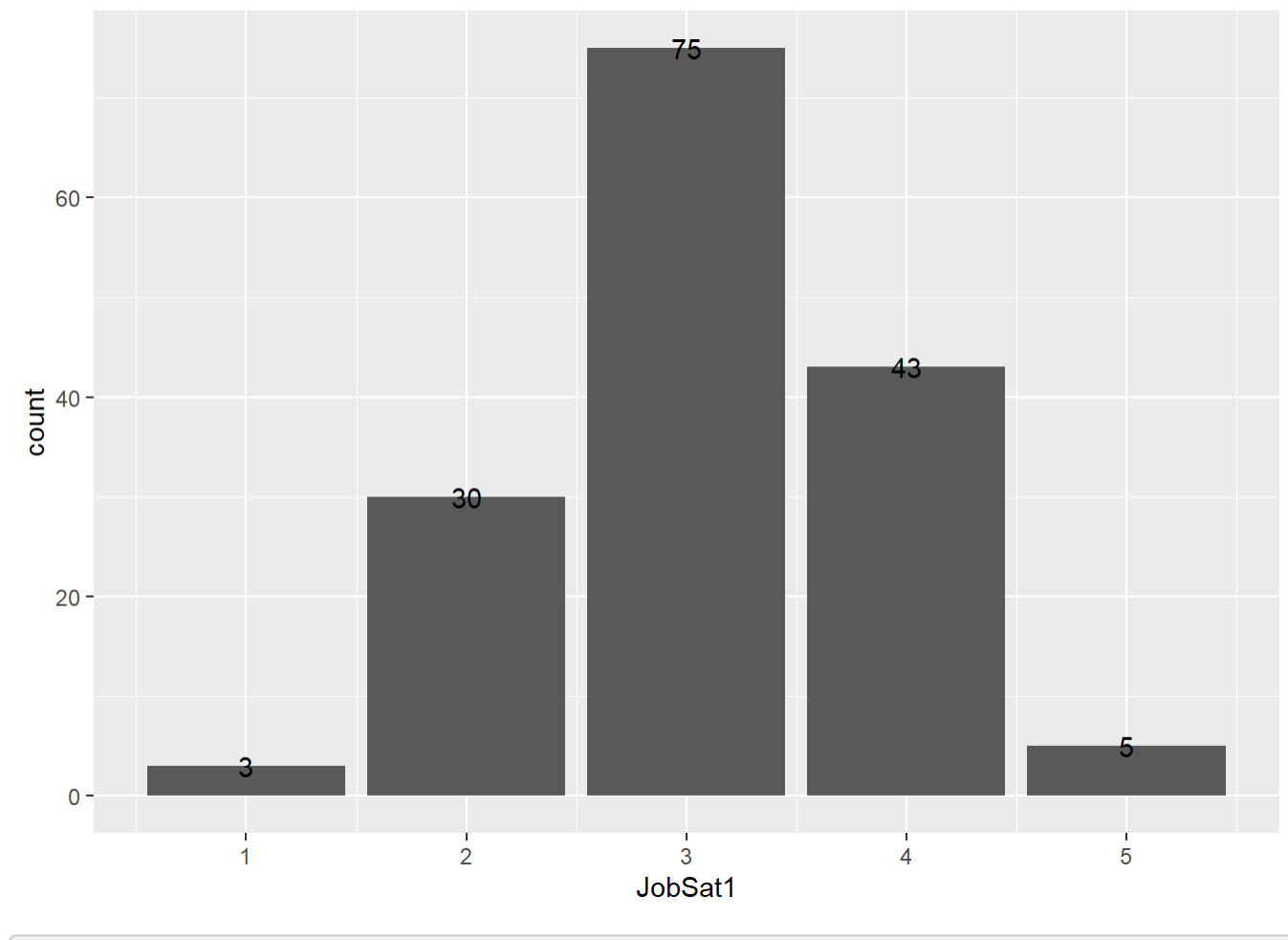
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



#Step-6: Data Visualization using ggplot2

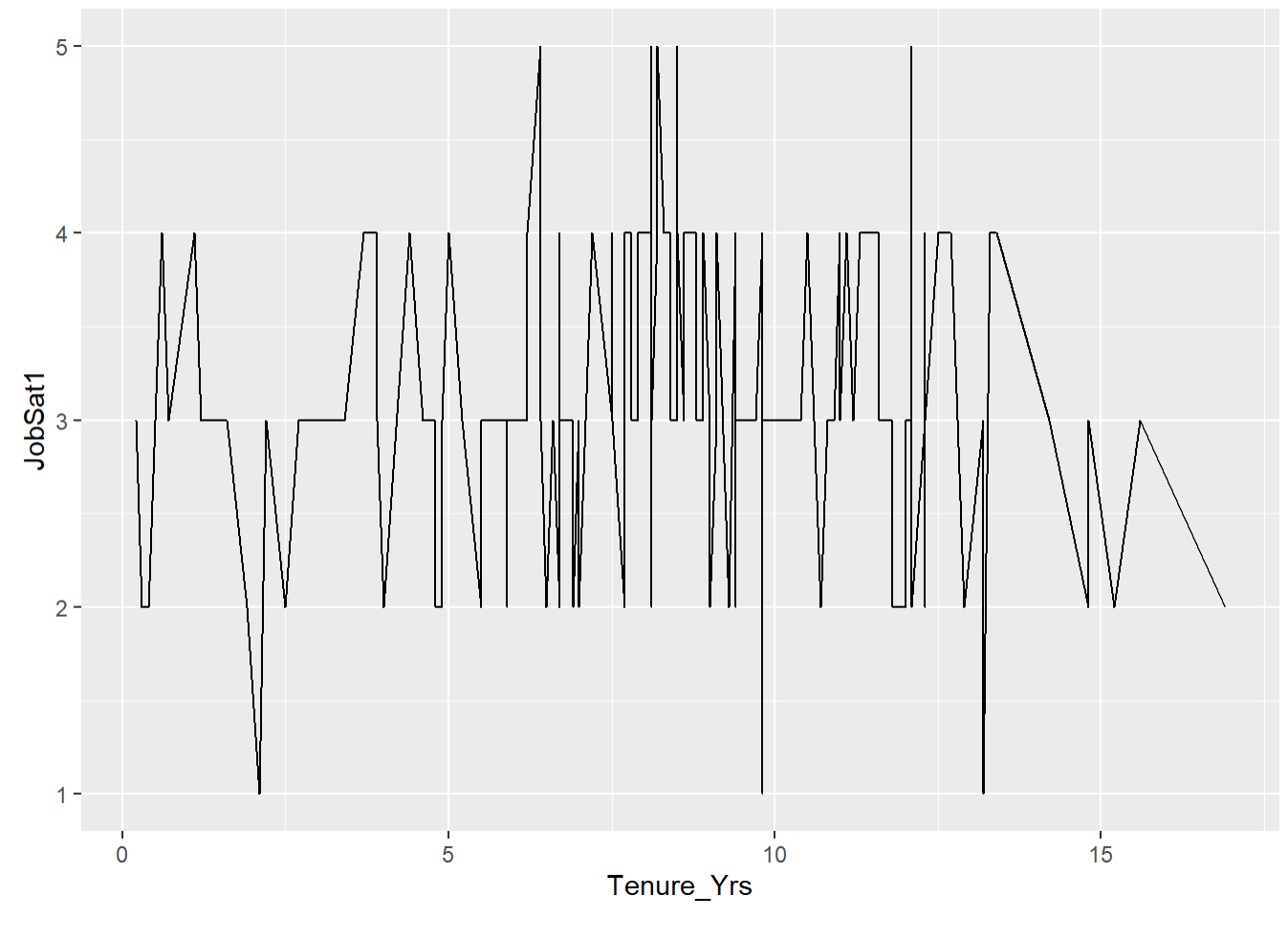
```
#Step-6a: Visualized Quantitative variable (JobSat1)

ggplot(surveydata, aes(x = JobSat1)) + geom_bar(stat = "count") + stat_count(geom = "text", aes(label = after_sta
t(count)), na.rm = TRUE)
```



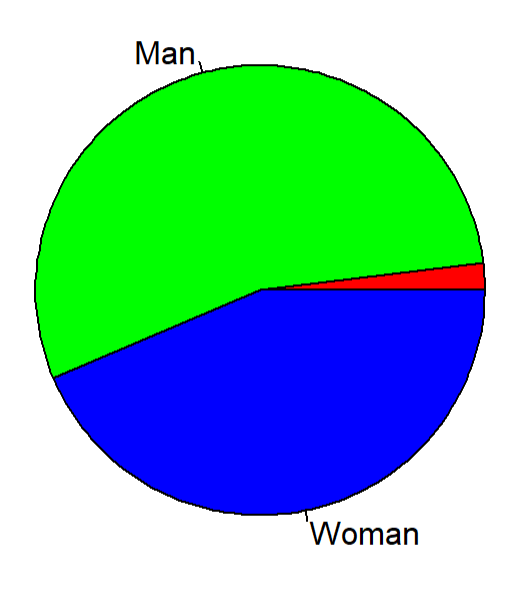

```
ggplot(data = surveydata, aes(x = Tenure_Yrs, y = JobSat1)) + geom_line()
```

```
## Warning: Removed 4 rows containing missing values (`geom_line()`).
```



#OBSERVATION FROM ABOVE LINE GRAPH: #The above bar graph illustrates the highest Job Satisfaction level lies at 3 across all the employees

```
#Step-6b: Visualized Categorical Variable (Gender)
#install.packages("plotrix")
library(plotrix)
pie(table(surveydata$Gender), labels = names(table(surveydata$Gender)), col = rainbow(3))
```




```
#Describe the proportions of the Gender (Male/Female)
prop.table(table(surveydata$Gender))*100
```

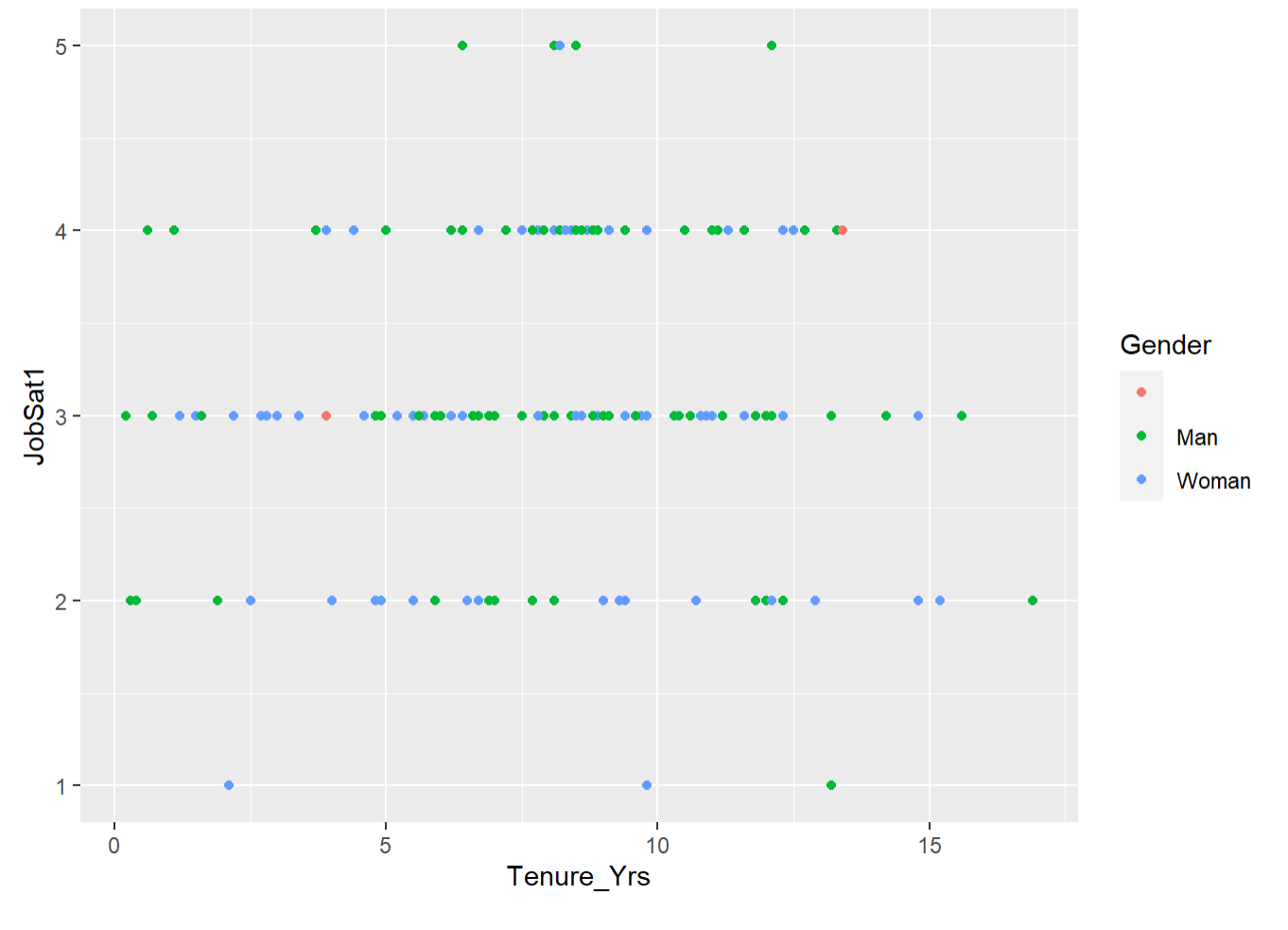
```
##
##                 Man     Woman
##  1.923077 54.487179 43.589744
```

#OBSERVATIONS FROM THE ABOVE PIECHART: #Majority of the proportion i.e., 54.4% for Man #Woman propotion is 43.5% #Nulls contributed to 1.9%

#Step-6c: Plotting a Scatter Plot

```
library(ggplot2) # loaded ggplot2 library to plot a scatterplot
ggplot(surveydata, aes(x = Tenure_Yrs, y = JobSat1, color = Gender)) + geom_point()
```

```
## Warning: Removed 4 rows containing missing values (`geom_point()`).
```



#OBSERVATIONS FROM THE ABOVE SCATTERPLOT: #The Scatter plot illustrates the Job Satisfaction level across different tenure for both Man and Woman.

#It clearly depicts that majority of the Man and Woman employees between tenure of 5-10 years are moderately satisfied (scale3 on Y-axis) and satisfied (scale4 on X-axis) with their job.