Univariate Analysis on nominal, ordinal, and interval/ratio variables

Ms. Nishigandha Wankhade

Dataset used: "adult.data" [1]

```
# Read the .data file

> data <- read.table("C:\\Users\\wankh\\Desktop\\Datasets\\adult\\adult.data", header = TRUE)

#Export the data to a CSV file

> write.csv(data, "C:\\Users\\wankh\\Desktop\\Datasets\\adult\\adult.csv", row.names = TRUE)

# To read the csv file to DATA FRAME

> df <- read.csv("C:\\Users\\wankh\\Desktop\\Datasets\\adult\\adult.csv", check.names = TRUE)

> view(df)
```

1. **Univariate analysis for "age" attribute (The Interval variable):**

A) Positions or Locations of "age" attribute:

I) Central Tendency:

i) Mean:

```
> mean(df$age)

[1] 38.58165
```

ii) Median:

```
> median(df$age)

[1] 37
```

iii) Mode:

#---- user define function for calculating mode () for age variable: It returns the most frequently occurring observation in the "adult.csv" dataset.

```
> age_vec <- c(df$age)
> mode_age <- Mode(age_vec)
>calculate_mode <- function(x) {
+   uniq_values <- unique(x)
+   uniq_cnt <- table(x)
+   mode_value <- uniq_values[which.max(uniq_cnt)]
+   return(mode_value)
+ }
> mode_age <- calculate_mode(age_vec)
> print(mode_age)
[1] 36
```

II) Quartiles, Percentiles and Deciles:

```
#------------Quartiles, percentiles and Deciles by age------------
> quantile(df[['age']], p = c(0, 0.25, 0.5, 0.75, 1))
  0%  25%  50%  75% 100%
  17   28   37   48   90
> quantile(df$age, probs = c(0.125, 0.375, 0.625, 0.875))  #probs(probability) to set various percentages
12.5% 37.5% 62.5% 87.5%
 23    32    42    56
```

```
#---------Deciles by age-------------
>  des_age <- quantile(df$age, probs = seq(0.1, 1, by = 0.1))
> print(des_age)
 10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
 22   26   30   33   37   41   45   50   58   90
```

B) Dispersion or Variability:

i) Range: is the difference between the highest and the lowest values a particular observation of the dataset

```
> range_value <- max(df$age) - min(df$age)
> print(range_value)
[1] 73
```

ii) Standard Deviation:

```
> sd(df$age)
[1] 13.64043
```

iii) Standard Error: is the way to measure how spread-out values are around them, in the dataset. The larger the standard error of the mean, the more spread-out values are around the mean in the dataset. And the sample size increases, it tends to decrease. [2]

```
#-------standard error-----() sd/ square root of n)------------
> print(sd(df$age)/sqrt(length(df$age)))
[1] 0.0755926
```

iv) Variance :

```
> var(df$age)
[1] 186.0614
```

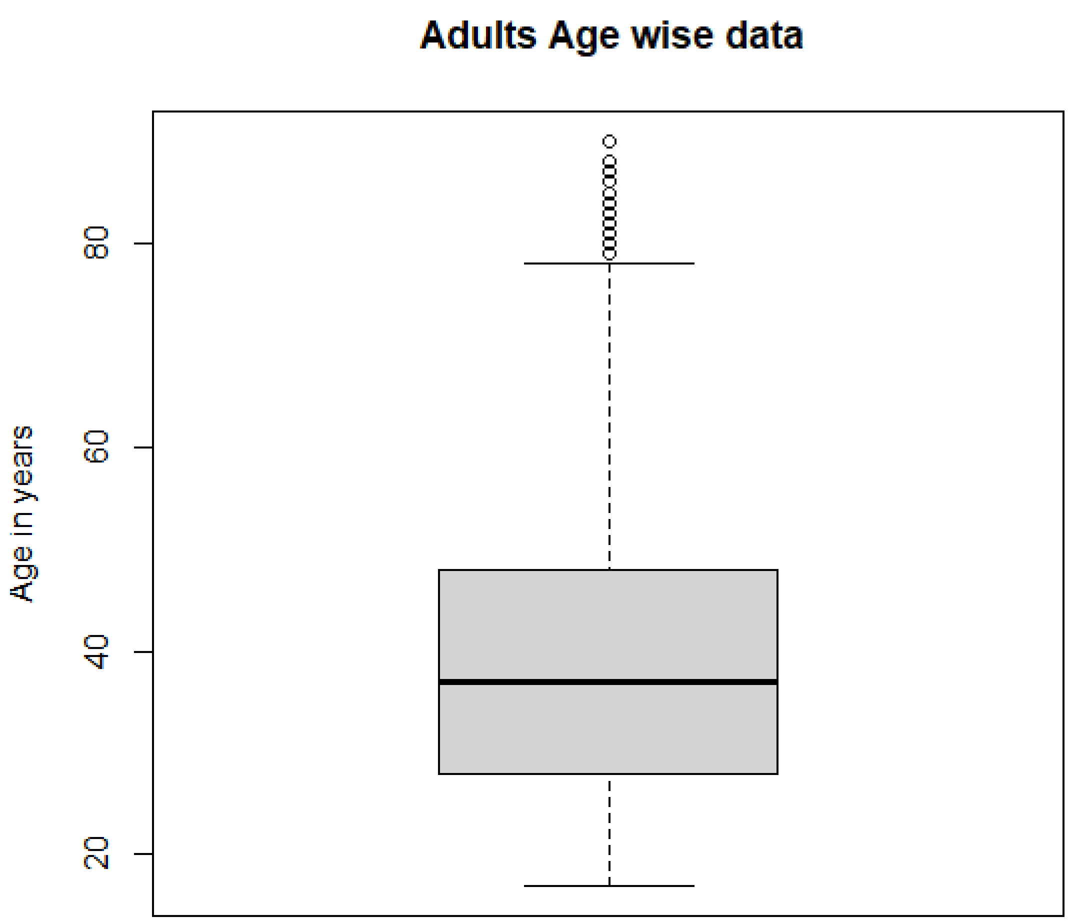
C) Graphs and Charts for “age” attribute:

i) Boxplot:

```
boxplot(df$age, main = " Adults Age wise data", ylab = "Age in years")
```

Figure 1

Box plot for adults having age range.

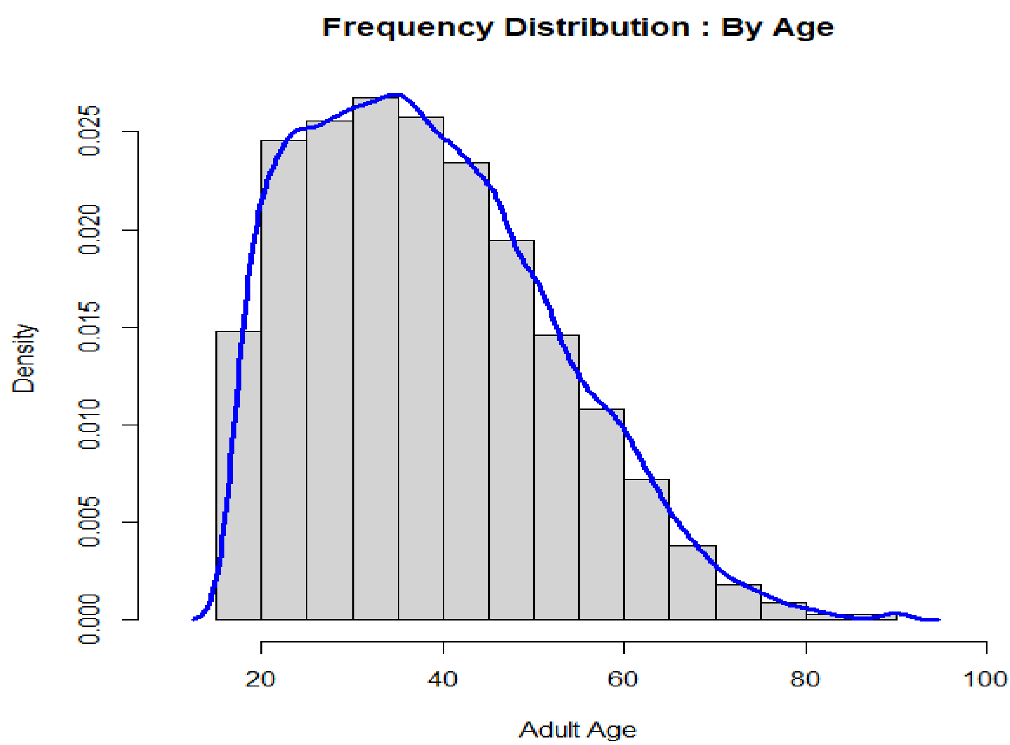


ii) Histogram by density:

```
> hist(age,
+       main = "Frequency Distribution : By Age",
+       xlab = "Adult Age",
+       xlim = c(10, 100),
+       freq = FALSE)
> lines(density(df$age), lwd=3, col= 'blue')
```

Figure 2

Histogram of density distribution by age.

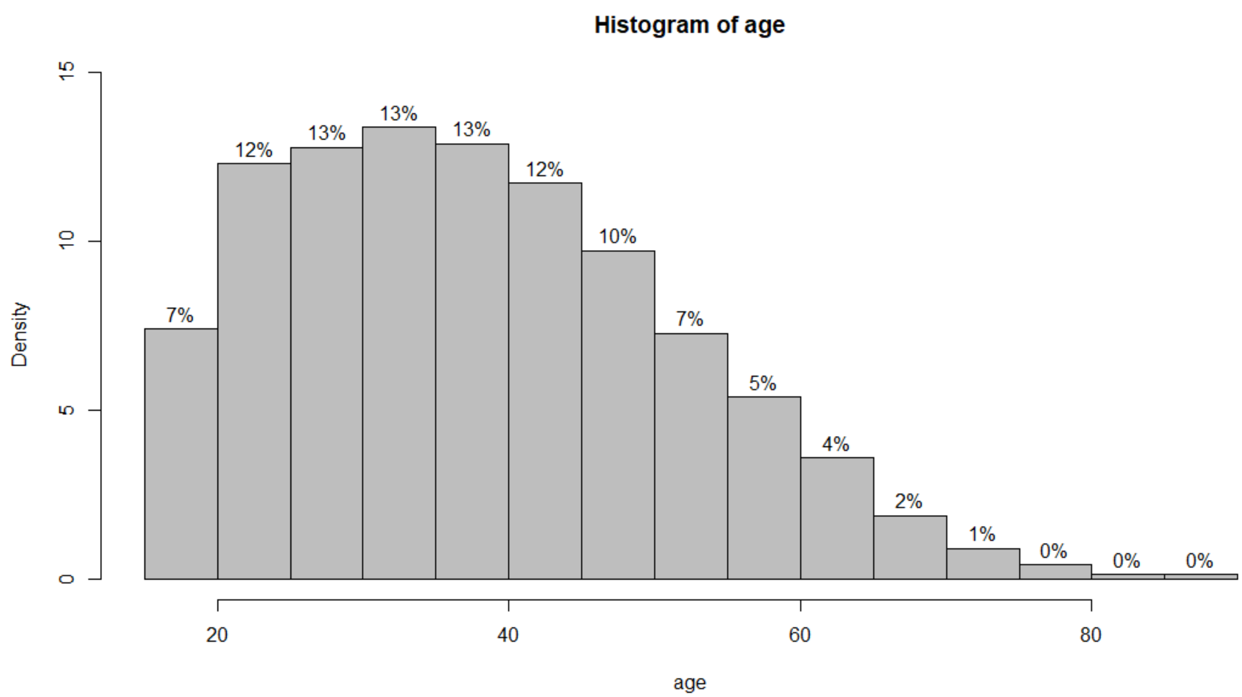


iii) Histogram by percentage:

```
> histPercent <- function(x, ...) {
+   H <- hist(age, plot = FALSE)
+   H$density <- with(H, 100 * density* diff(breaks)[1])
+   labs <- paste(round(H$density), "%", sep="")
+   plot(H, freq = FALSE, labels = labs, ylim=c(0, 1.08*max(H$density)),...)
+ }
> histPercent(df$age, col="gray")
```

Figure 3

Histogram to represent "age" distribution by percentage.

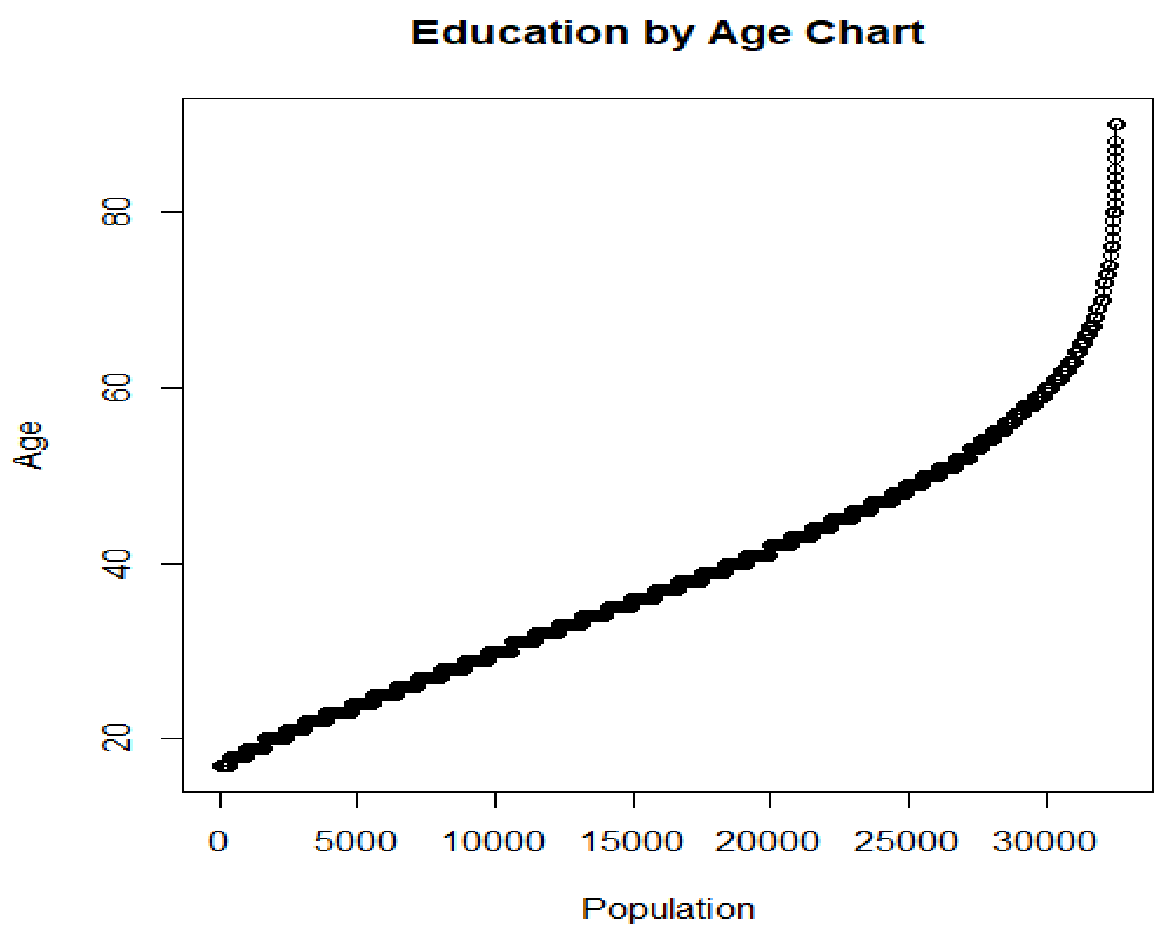


iv) Line plot:

```
plot(df$age, type = "o",
+       xlab = "Population", ylab = "Age",
+       main = "Education by Age Chart")
```

Figure 4

Line plot for all population by age..

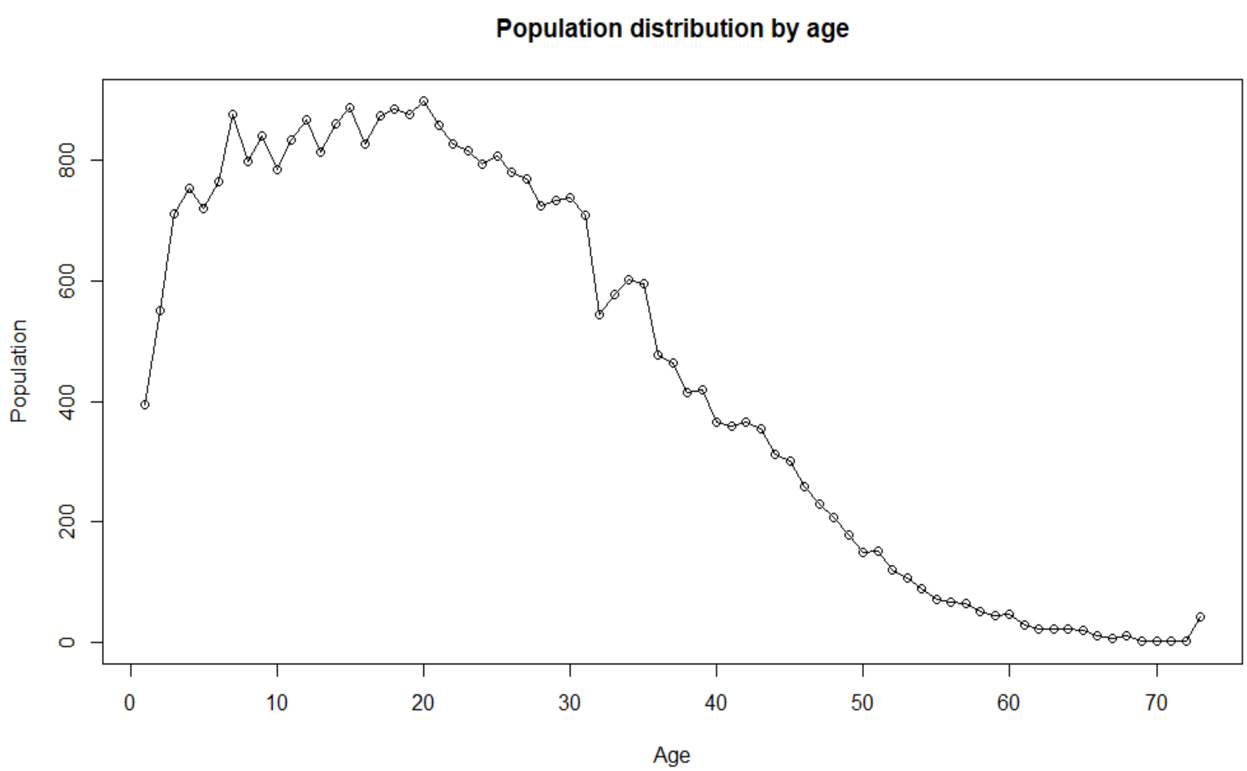


v) Line chart:

```
#------------Line chart by age-----------
df_age <- data.frame(freq_table_age)
plot(df_age$Freq, type = "o",
xlab = "Age", ylab = "Population",
main = "Population distribution by age")
```

Figure 5

Line chart for all population by age.



vi) Stem and Leaf plot: Is a textual graph, used to classify.

stem(df$age)

The decimal point is 1 digit(s) to the right of the |

1 | 77777777777777777777777777777777777777777777777777777777777777777777777777+1577

2 | 00000000000000000000000000000000000000000000000000000000000000000000000000+3833

2 | 55555555555555555555555555555555555555555555555555555555555555555555555555+4061

3 | 00000000000000000000000000000000000000000000000000000000000000000000000000+4258

3 | 55555555555555555555555555555555555555555555555555555555555555555555555555+41954 |

00000000000000000000000000000000000000000000000000000000000000000000000000+3796

4 | 55555555555555555555555555555555555555555555555555555555555555555555555555+32195 |

00000000000000000000000000000000000000000000000000000000000000000000000000+2474

5 | 55555555555555555555555555555555555555555555555555555555555555555555555555+17846 |

00000000000000000000000000000000000000000000000000000000000000000000000000+1228 6 | 5

5555555555555555555555555555555555555555555555555555555555555555555555555+627 7 | 000

00000000000000000000000000000000000000000000000000000000000000000000000+26

7 | 55555555555555555555555555555555555555555555555555566666666666666666666666+85

8 | 00000000000000000000001111111111111111111111111122222222222233333344444444

8 | 55567888

9 | 00000000000000000000000000000000000000000000000

D) Frequency Distribution for "age" (Interval Variable):

i) Average frequency and Cumulative frequency:

```
freq_table_age<- table(df$age)    # table() generates frequency table
> set.seed(1)
> cumfreq_age <- cumsum(freq_table_age)      # cumsum() for cumulative frequencies
> data_frame <- data.frame(freq_table_age, cumfreq_age)
> colnames(data_frame) <- c("Age", "Frequency", "Cumulative_Frequency")
> print(data_frame)
```

	Age	Frequency	Cumulative_Frequency
17	17	395	395
18	18	550	945
19	19	712	1657
20	20	753	2410
21	21	720	3130
22	22	765	3895
23	23	877	4772
…		…	…
87	87	1	32515
88	88	3	32518
90	90	43	32561

2. Univariate analysis for "occupation" attribute (Nominal variable):

I) Frequency table:

```
> freq_table_occu <- table(df$occupation)
> print(freq_table_occu)
```

?	Adm-clerical	Armed-Forces	Craft-repair	Exec-managerial
1843	3770	9	4099	4066

Farming-fishing	Handlers-cleaners	Machine-op-inspct	Other-service	Priv-house-serv
994	1370	2002	3295	149

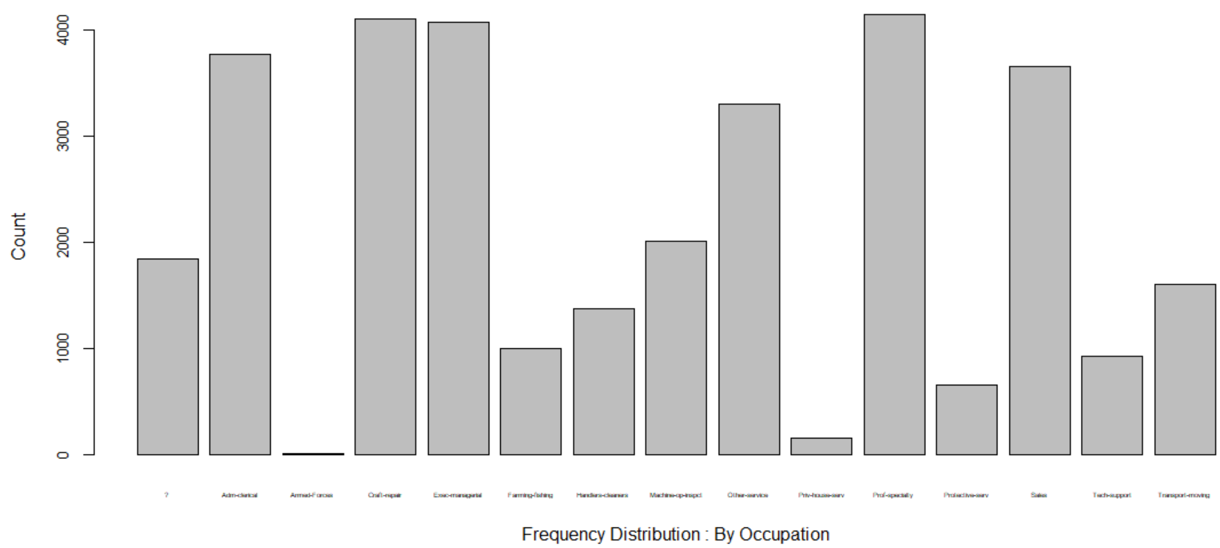
Prof-specialty	Protective-serv	Sales	Tech-support	Transport-moving
4140	649	3650	928	1597

II) Bar chart:

```
> barplot(table(df$occupation), cex.axis = 0.8, cex.names = 0.4,
+       xlab = 'Frequency Distribution : By Occupation',
+       ylab = 'Count')
```

Figure 6

Bar chart by occupation.



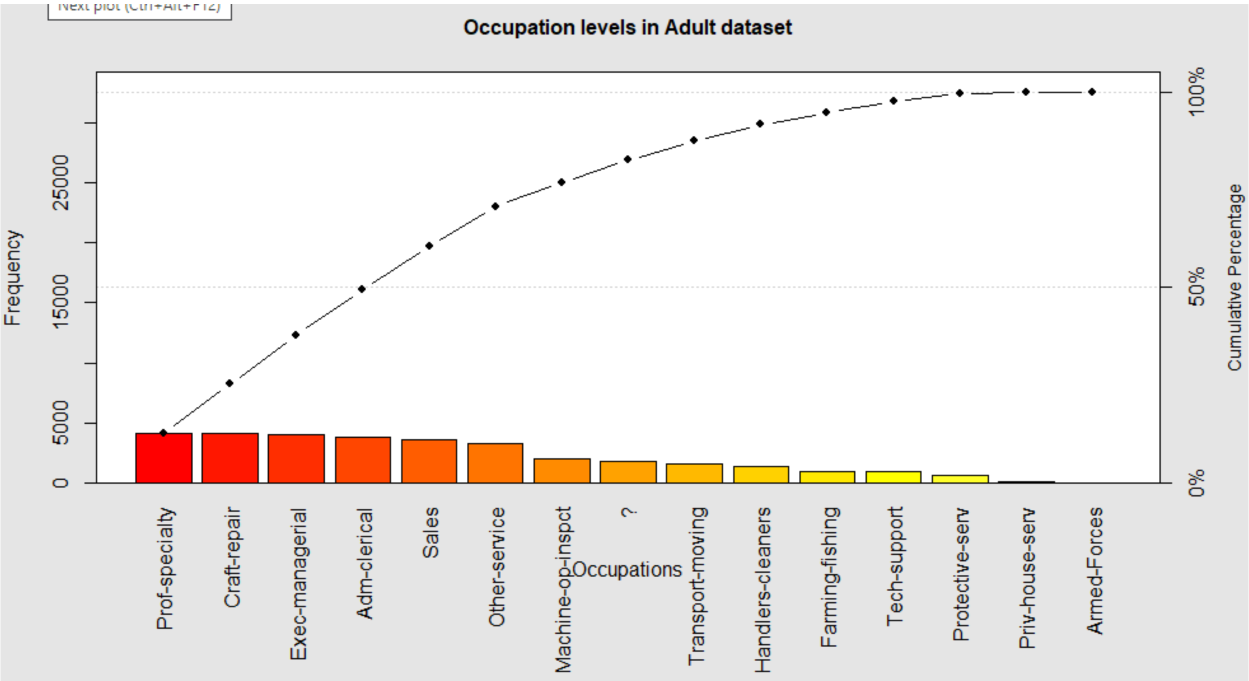
III) Pareto Chart: It is used to show the frequency of occurrences of the event in different categories in decreasing order, and an overlaid line chart indicates the cumulative percentage of occurrences.

```
install.packages('qcc')
library(qcc)
defect_occu <- c(1843,3770,9,4099,4066,994,1370,2002,3295,149,4140,649,3650,928,1597)
#x axis titles
names(defect_occu) <- c("?","Adm-clerical", "Armed-Forces", "Craft-repair",
```

```
        "Exec-managerial", "Farming-fishing", "Handlers-cleaners", "Machine-op-inspct",
        "Other-service", "Priv-house-serv","Prof-specialty","Protective-serv",
        "Sales","Tech-support", "Transport-moving")
    pareto.chart(defect_occu, xlab = "Occupations",
        ylab = "Frequency",
        col=heat.colors(length(df_occu)),
        cumperc = seq(0, 1000, by = 50),
        ylab2 = "Cumulative Percentage", #lable y right
        main = "Occupation levels in Adult dataset", #title of the chart)
```

Figure 7

Pareto chart by occupation.



Pareto chart analysis for defect_occu

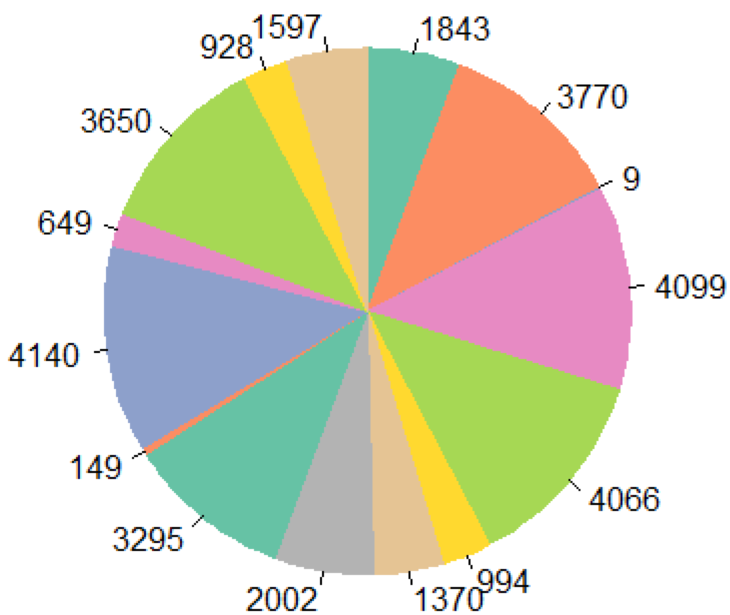
	Frequency	Cum.Freq.	Percentage	Cum.Percent.
Prof-specialty	4.140000e+03	4.140000e+03	1.271460e+01	1.271460e+01
Craft-repair	4.099000e+03	8.239000e+03	1.258868e+01	2.530328e+01
Exec-managerial	4.066000e+03	1.230500e+04	1.248733e+01	3.779061e+01
Adm-clerical	3.770000e+03	1.607500e+04	1.157827e+01	4.936888e+01
Sales	3.650000e+03	1.972500e+04	1.120973e+01	6.057861e+01
Other-service	3.295000e+03	2.302000e+04	1.011947e+01	7.069807e+01
Machine-op-inspct	2.002000e+03	2.502200e+04	6.148460e+00	7.684653e+01
?	1.843000e+03	2.686500e+04	5.660146e+00	8.250668e+01
Transport-moving	1.597000e+03	2.846200e+04	4.904641e+00	8.741132e+01
Handlers-cleaners	1.370000e+03	2.983200e+04	4.207487e+00	9.161881e+01
Farming-fishing	9.940000e+02	3.082600e+04	3.052732e+00	9.467154e+01
Tech-support	9.280000e+02	3.175400e+04	2.850035e+00	9.752157e+01
Protective-serv	6.490000e+02	3.240300e+04	1.993182e+00	9.951476e+01
Priv-house-serv	1.490000e+02	3.255200e+04	4.576027e-01	9.997236e+01
Armed-Forces	9.000000e+00	3.256100e+04	2.764043e-02	1.000000e+02

IV) Pie chart:

```
library(RColorBrewer)

install.packages("plotrix")

library(plotrix)

occu_count <- c(dataFrame_occu$Freq)

color <- brewer.pal(length(occu_count), "Set2")

pie(occu_count, clockwise = TRUE, labels = occu_count, col = color, cex = 1, border = color)
```

Figure 8

Pie chart by occupation.



3. Univariate analysis for "education.num" attribute (Ordinal variable):

A) Positions or Location of "education.num" attribute:

I) Central Tendency:

i) > mean(df$education.num)

[1] 10.08068

ii) > median(df$education.num)

[1] 10

iii) > summary(df$education.num)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0    9.00   10.00   10.08   12.00   16.00
```

B) Dispersion or Variability for "education.num" attribute:

i) Standard deviation:

> sd(df$education.num)

[1] 2.57272

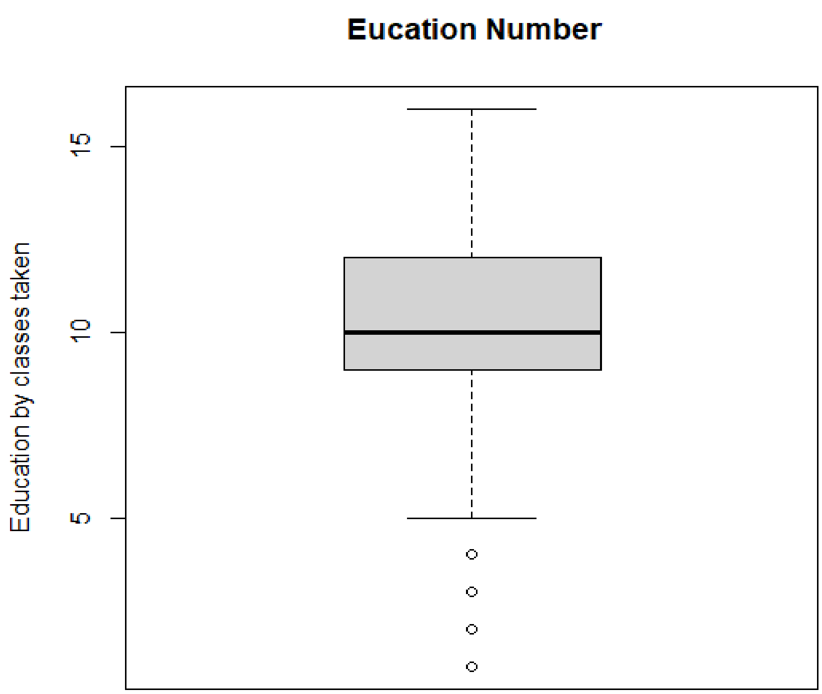
C) Graphs and Charts for "education.num" attribute:

i) Boxplot:

>boxplot(df$education.num, main = " Eucation Number", ylab = "Education by class
es taken")

Figure 9

Box plot for education number.

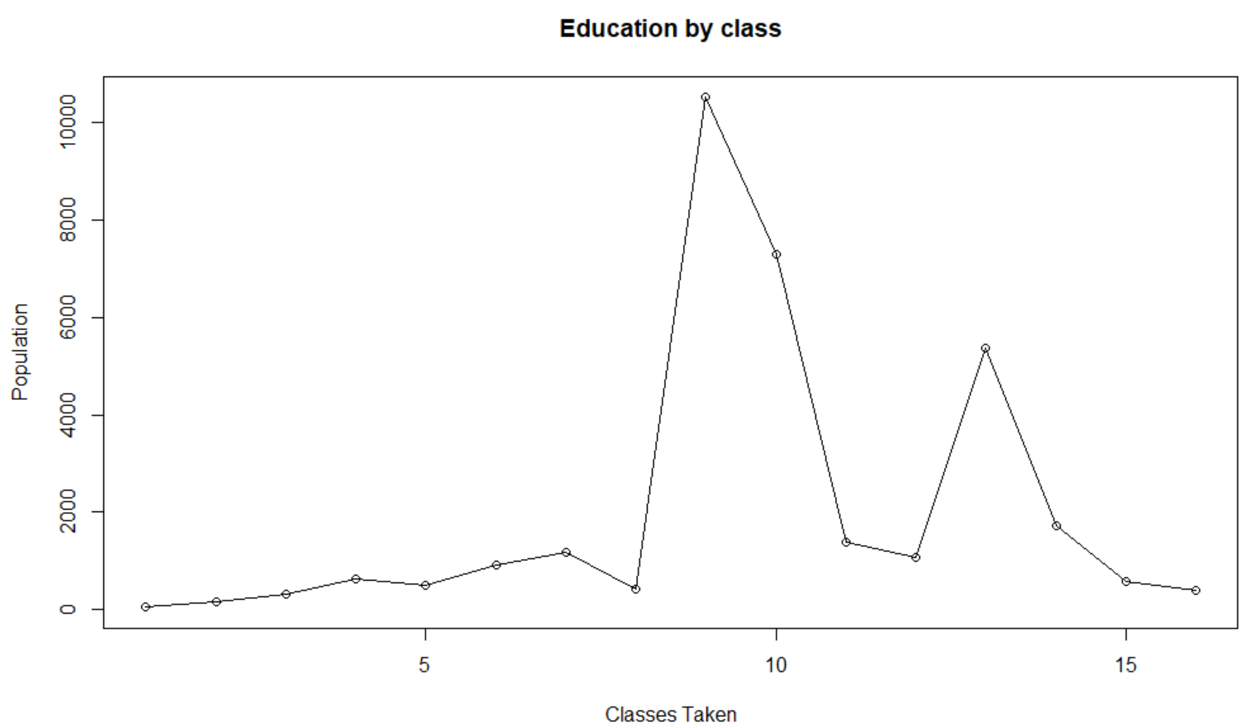


ii) Line chart:

```
> freq_eduNum <- table(df$education.num)

> df_eduNum <- data.frame(freq_eduNum)

> plot(df_eduNum$Freq, type = "o",  xlab = "Classes Taken", ylab = "Population",
 main = "Education by class")
```

Figure 10

Line chart by education number.

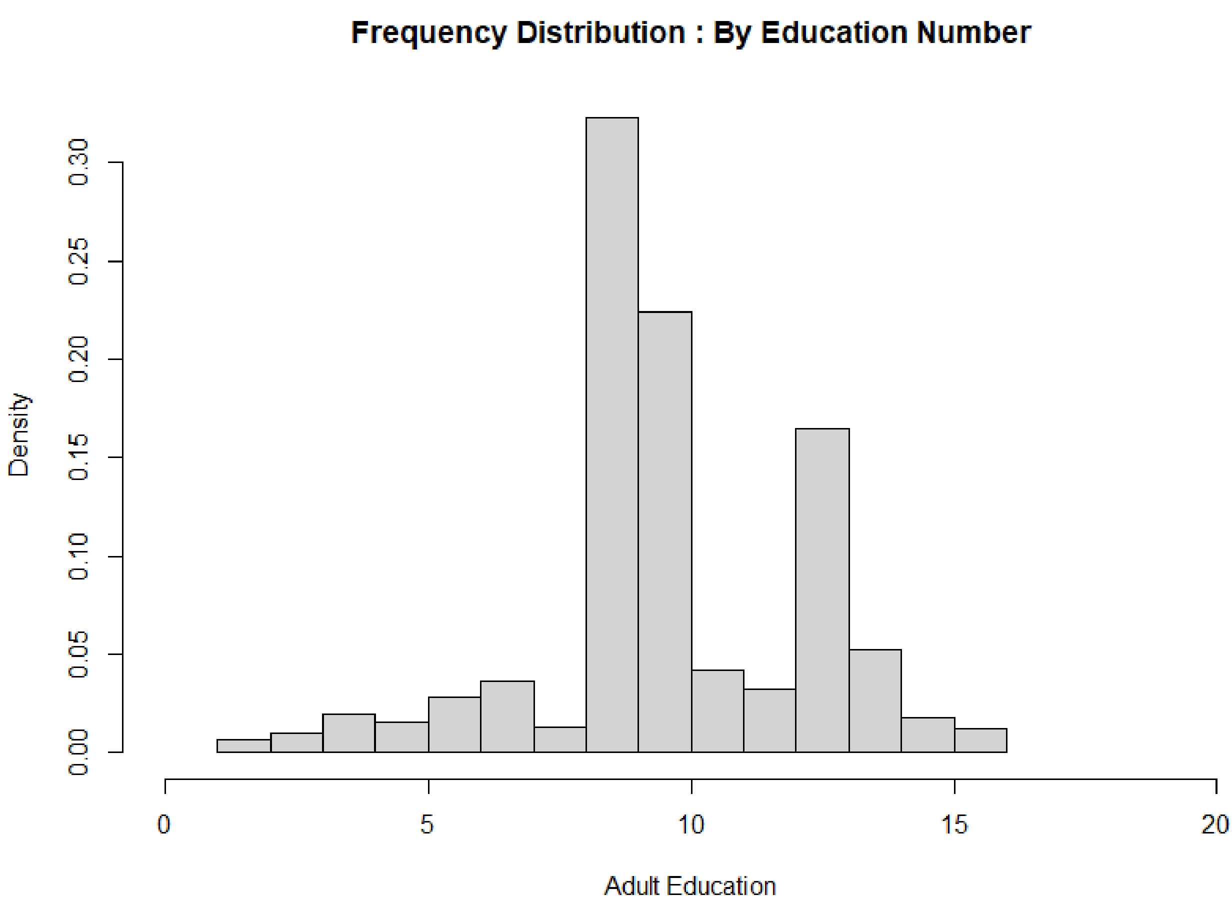


iii) Histogram :

```
> hist(df$education.num,
+     main = "Frequency Distribution : By Education Number",
+     xlab = "Adult Education",
+     xlim = c(0, 20),
+     freq = FALSE)
```

Figure 11

Histogram of education number.

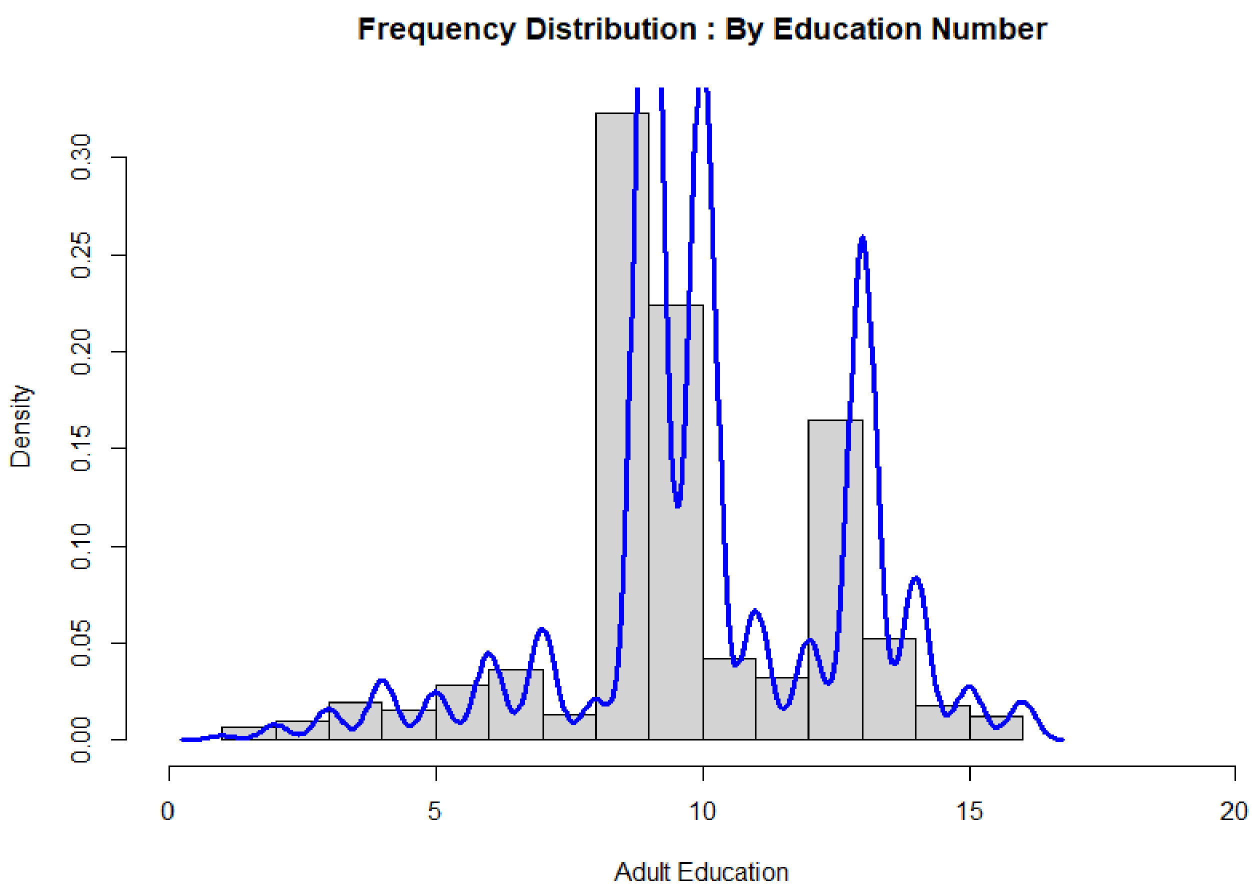


iv) Histogram by density:

> > lines(density(df$education.num), lwd=3, col= 'blue')

Figure 12

Density distribution by education number.



4. Univariate analysis for other nominal attributes: (EXTRAS)

a) "workclass" nominal attribute:

i) Frequency table:

```
> freq_table_wc<- table(df$workclass)
> #print(freq_table_wc, row.names = TRUE)
> cbind(freq_table_wc)
 freq_table_wc
?                  1836
Federal-gov         960
Local-gov          2093
Never-worked          7
Private           22696
Self-emp-inc       1116
Self-emp-not-inc   2541
State-gov          1298
Without-pay          14
```

ii) Bar Chart:

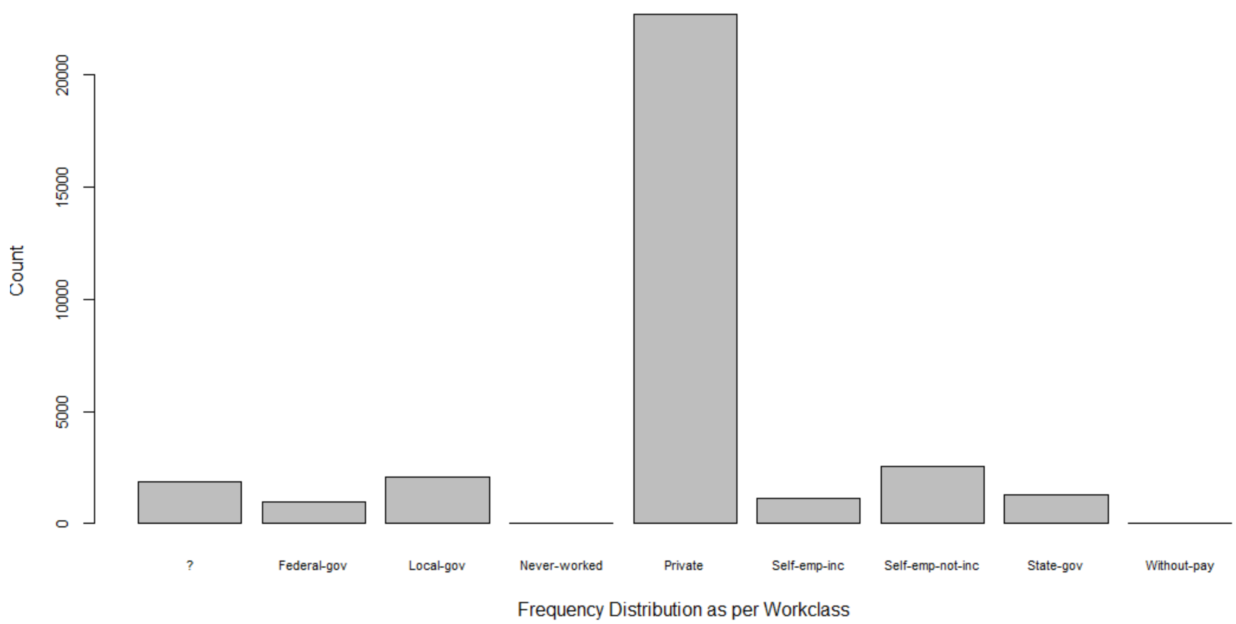
barplot(table(df$workclass), cex.axis = 0.8, cex.names = 0.7,

\+ xlab = 'Frequency Distribution as per Workclass',

\+ ylab = 'Count')

Figure 13

Box plot for workclass.



b) "sex" nominal attribute:

i) Frequency table:

freq <- table(df$sex)

> print("Frequency count of column SEX")

> print(freq)

[1] "Frequency count of column SEX"

Female Male

10771 21790

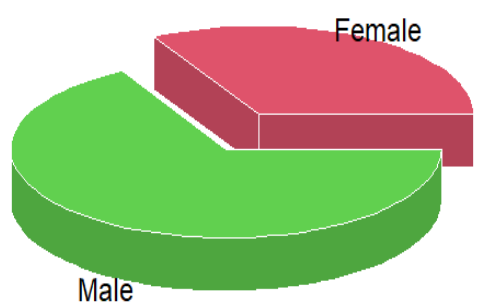
```
#---------------create freq table by group using dplyr pkg----------------

install.packages('dplyr')

library(dplyr)

df %>%

+   group_by(df$sex, df$age) %>%

+   summarize(Freq=n())   #`summarise()` has grouped output by 'df$sex'.

# A tibble: 144 × 3

# Groups:   df$sex [2]

  `df$sex` `df$age`  Freq

  <chr>       <int> <int>

 1 Female        17   186

 2 Female        18   268

 3 Female        19   356

 4 Female        20   363

 5 Female        21   329

 6 Female        22   342

 7 Female        23   359

 8 Female        24   305

 9 Female        25   313

10 Female        26   290

# ℹ 134 more rows
```

ii) Pie chart:

```
> install.packages("lessR")
> library(lessR)
> install.packages("plotrix")
> library(plotrix)
> gender_count <- table(df$sex)
> df_gender <- data.frame(gender_count)
> pie3D(df_gender$Freq, labels = df_gender$Var1, explode = 0.15, col = 2:3, label
        col = "black", border = "white")
```

Figure 14

Pie chart for gender.

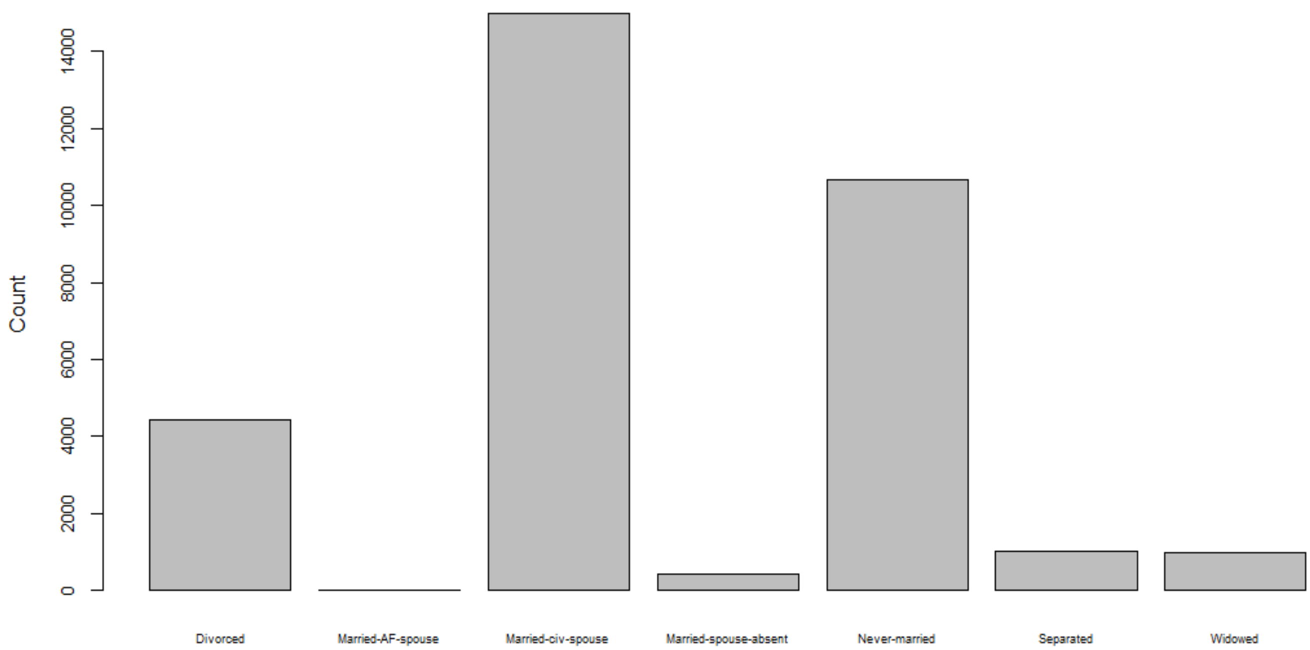


c) "marital-status" nominal attribute:

```
> barplot(table(df$marital.status), cex.axis = 0.8, cex.names = 0.55,
+       xlab = 'Frequency Distribution : By Marital Status',
+       ylab = 'Count')
```

Figure 15

Box plot for marital status.



How does each data type influence the kind of analysis and insights?

The types of data (nominal, ordinal interval/ratio) significantly identify the appropriate statistical analysis to be performed on the dataset.

- Nominal data: appropriate for categorical analysis (frequency distribution by category, mode, bar chart, pie chart, Pareto chart), non-parametric tests.

- Ordinal data: appropriate for rank-based statistics and non-parametric tests
- Interval data: appropriate for a range of parametric statistics like:
  - to identify position and location
    - Central tendency (mean, mode, median)
    - Quantiles (quartiles, deciles, percentiles)
  - to find dispersion or variability
    - range, average, variance, standard deviation, standard error, and coefficien t of variance
  - to find the shape of attributes
    - skewness
    - kurtosis

Interpretation:

Based on univariate descriptive statistics performed on nominal, ordinal, and interval v ariables from the "adult" dataset, we reached the following conclusion:

The majority of males aged between 20 and 40, with an average age of 36-37 years, are either ma rried or never married. Their education levels range from 8^{th} – 11^{th} grade. Most of them work in t he private sector, holding occupations in professional specialty, craft-repair, or executive-manage rial positions.

Reflection:

It is important to understand the type of each data so that we can use appropriate statisti cal methods to analyze the given dataset. It will help us to ensure the accuracy of data representat ions and to effectively communicate the results, as not all audiences can understand technical det ails. So, graphical representations to explain what is happening with data help the audience to un derstand and predict future measures needed to be taken on the data, i.e., help in decision making . Misinterpretation of statistical methods can severely degrade the company's performance by im pacting research outcomes, policy decisions as well as business strategies.

References:

Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Bobbitt, Z. (2020, October 2). *How to Calculate Standard Error of the Mean in R*. Statology.org. Retrieved May 17, 2024, from https://www.statology.org/standard-error-of-mean-r/