Assignment 2

Giovanni De Francesco

2024-10-17

```
library(car)
```

```
## Caricamento del pacchetto richiesto: carData
```

```
data("Salaries")
```

Before starting to describe the data, it's important to check that the einvoriment is correctly reading it.

```
suppressWarnings(library(dplyr))
```

```
##
## Caricamento pacchetto: 'dplyr'

## Il seguente oggetto è mascherato da 'package:car':
##
##     recode

## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag

## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(Salaries)
```

```
## Rows: 397
## Columns: 6
## $ rank          <fct> Prof, Prof, AsstProf, Prof, Prof, AssocProf, Prof, Prof,~
## $ discipline    <fct> B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, A, A,~
## $ yrs.since.phd <int> 19, 20, 4, 45, 40, 6, 30, 45, 21, 18, 12, 7, 1, 2, 20, 1~
## $ yrs.service   <int> 18, 16, 3, 39, 41, 6, 23, 45, 20, 18, 8, 2, 1, 0, 18, 3,~
## $ sex           <fct> Male, Male, Male, Male, Male, Male, Male, Male, Male, Fe~
## $ salary        <int> 139750, 173200, 79750, 115000, 141500, 97000, 175000, 14~
```

Every variable of the dataset is correctly expressed. Now let's check if there are any missing values.

```r
sum(is.na(Salaries))
```

```
## [1] 0
```

Salaries doesn't present any missing values.

By using the summary function, we can have a general idea of the characteristics of the data. The result varies in respect to the type of variable, such as double, integer, factor etc.

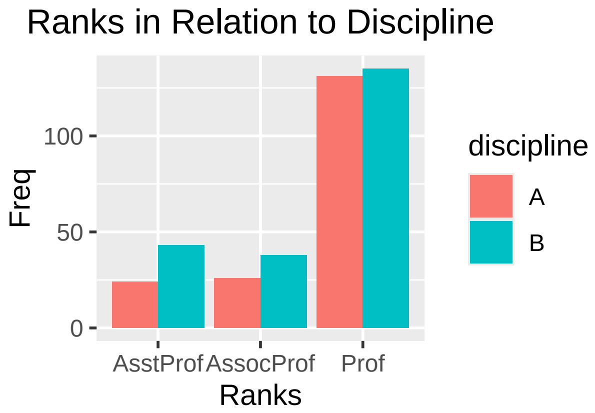
```r
summary(Salaries)
```

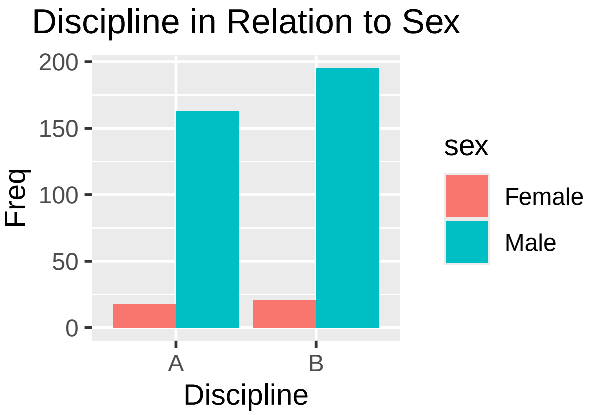
```
##         rank     discipline yrs.since.phd    yrs.service        sex     
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39  
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358  
##  Prof     :266              Median :21.00   Median :16.00               
##                             Mean   :22.31   Mean   :17.61               
##                             3rd Qu.:32.00   3rd Qu.:27.00               
##                             Max.   :56.00   Max.   :60.00               
##      salary      
##  Min.   : 57800  
##  1st Qu.: 91000  
##  Median :107300  
##  Mean   :113706  
##  3rd Qu.:134185  
##  Max.   :231545
```

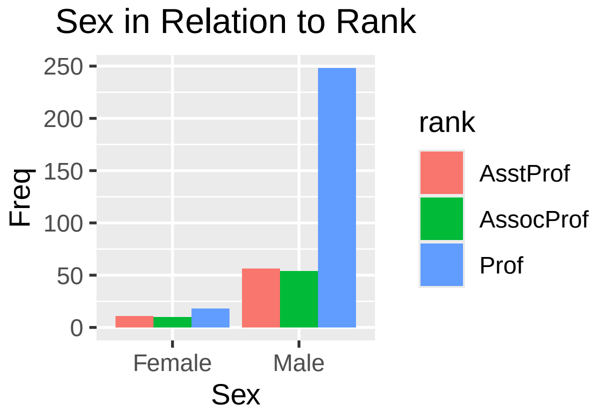
```r
library(ggplot2)
library(gridExtra)
```

```
## 
## Caricamento pacchetto: 'gridExtra'

## Il seguente oggetto è mascherato da 'package:dplyr':
## 
##     combine
```

```r
p1<-ggplot(Salaries,aes(x=rank,fill=discipline))+
  labs(x="Ranks",y="Freq",title = "Ranks in Relation to Discipline")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")
p2<-ggplot(Salaries,aes(x=discipline,fill=sex))+
  labs(x="Discipline",y="Freq",title = "Discipline in Relation to Sex")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")
p3<-ggplot(Salaries,aes(x=sex,fill=rank))+
  labs(x="Sex",y="Freq",title = "Sex in Relation to Rank")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")

grid.arrange(p1,p2,p3,ncol=2,nrow=2)
```

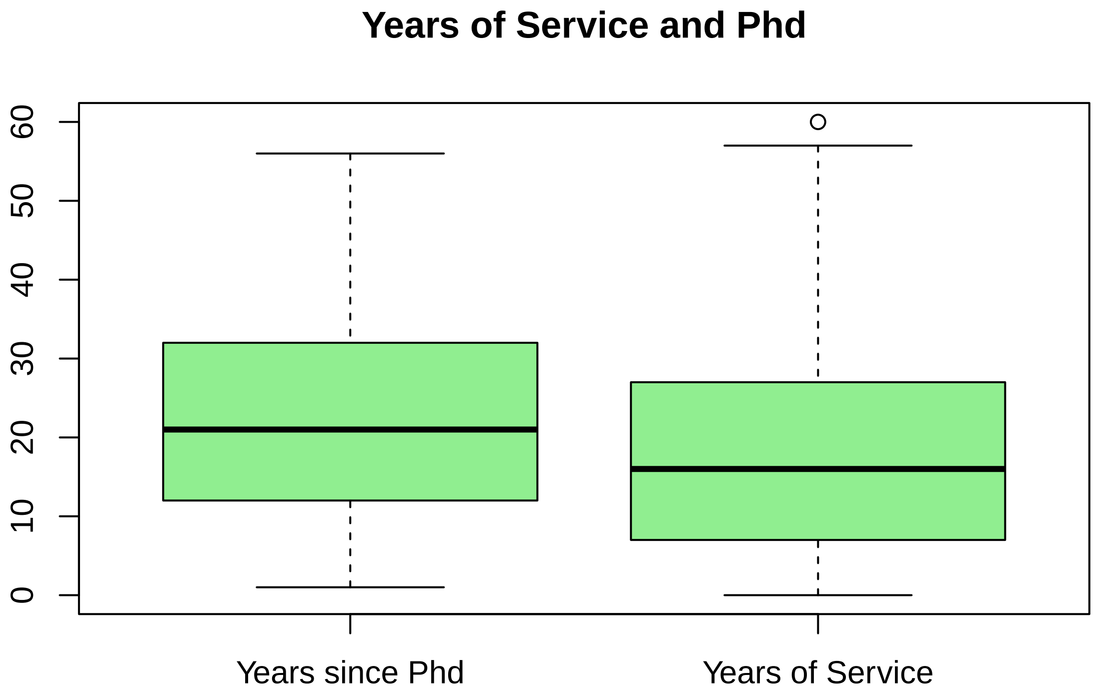

Ranks in Relation to Discipline


Discipline in Relation to Sex

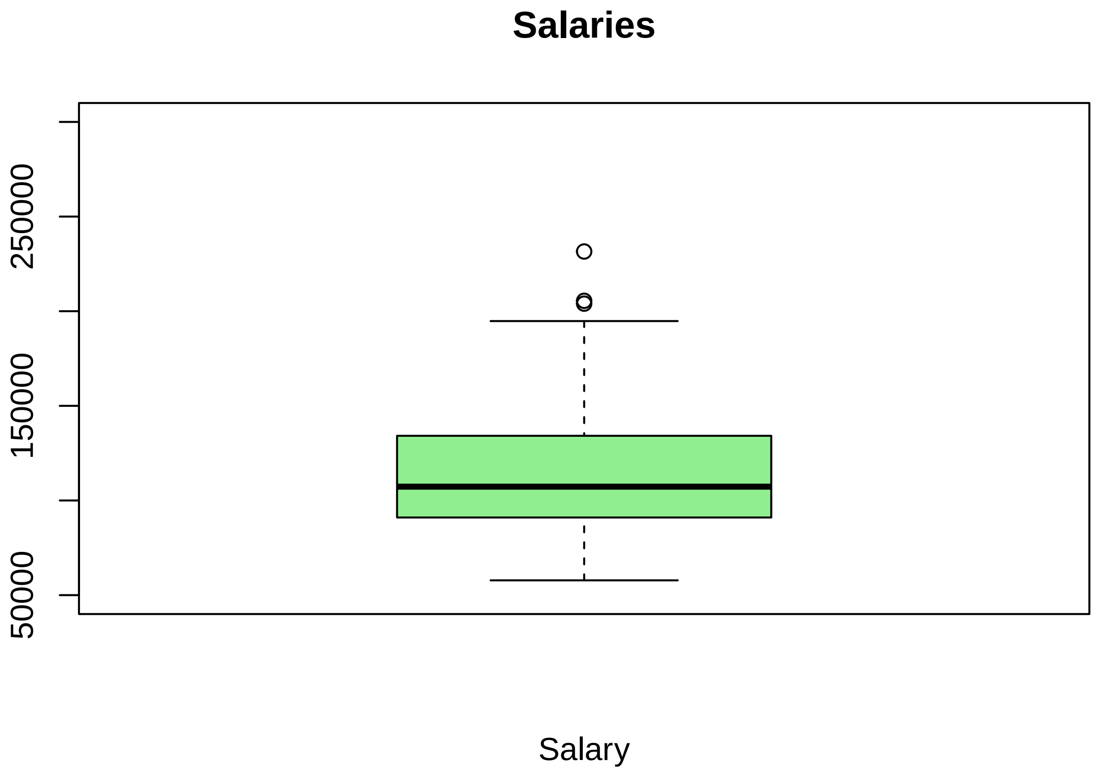

Sex in Relation to Rank

These charts provide insights into various relationships within the salaries dataset, highlighting aspects such as discipline, gender, and academic rank. The first chart shows how different academic ranks—Assistant Professor, Associate Professor, and Professor—are distributed across two disciplines, A and B. In Discipline B, there is a clear predominance of full Professors, while Discipline A has a more balanced distribution across ranks. The second chart explores the distribution of individuals across disciplines based on gender. Males represent the majority in both disciplines, with a particularly strong presence in Discipline B. The third chart examines the gender breakdown within each academic rank. Males are significantly overrepresented in the Professor rank, while females have lower representation across all ranks.

```
boxplot(Salaries$yrs.since.phd,Salaries$yrs.service,col = "lightgreen",names = c("Years since Phd","Yea
title("Years of Service and Phd")
```



```
boxplot(Salaries$salary,col="lightgreen",xlab="Salary",ylim=c(50000,300000))
title("Salaries")
```



The range (whiskers) for “Years since PhD” goes approximately from 0 to 60, while for “Years of Service,” it goes from about 0 to 50. Both distributions seem to have a similar interquartile range (IQR), indicating that both variables have similar variability. “Years of Service” has a slightly lower median compared to “Years since PhD.” There are no apparent outliers in either category.

The other boxplot represents the distribution of salaries. The salary values range from approximately 55000 to a bit over 200000, with the IQR showing most salaries between 100000 and 150000. The median salary appears to be just below 150000. There are a couple of outliers on the higher end, around 200000.

These boxplots give a quick visual comparison of “Years since PhD,” “Years of Service,” and “Salaries”.

–Gender Gap Pay–

Analyzing the presence of the gender gap pay can be very tricky. Data must be strictly studied in order to avoid any potential inequality which would mislead the results.

At first sight two potential issues appear:

1) There is an enormous inequality in the representation of male and females.

```
round(prop.table(table(Salaries$sex))*100,3)
```

```
##
## Female   Male
##  9.824 90.176
```

2) Professors, Associates, and Assistants may have very different salaries; if there is a significant difference in the representation of males and females across ranks, analyzing the dataset without accounting for these divisions can lead to misleading results.

```r
library(dplyr)

Ass<-Salaries %>%
  filter(rank=="AsstProf")

Assoc<-Salaries %>%
  filter(rank=="AssocProf")

Prof<-Salaries %>%
  filter(rank=="Prof")

mean(Ass$salary)
```

```
## [1] 80775.99
```

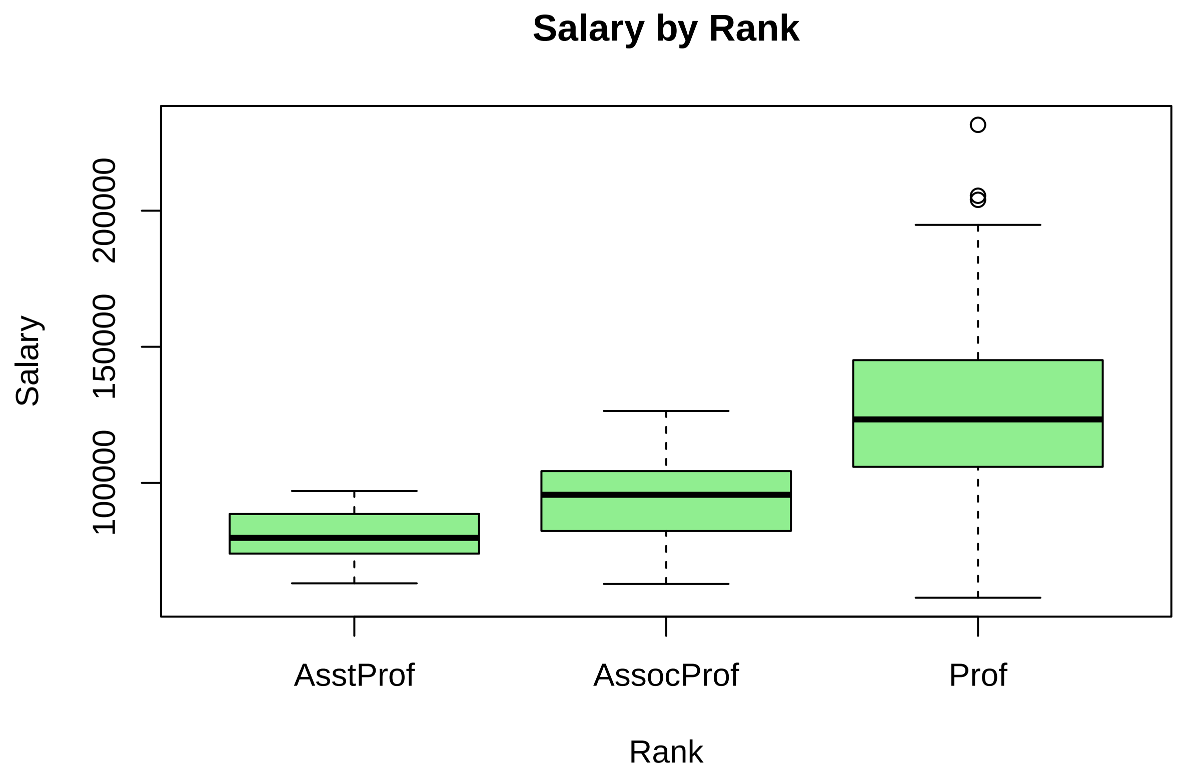
```r
mean(Assoc$salary)
```

```
## [1] 93876.44
```

```r
mean(Prof$salary)
```

```
## [1] 126772.1
```

```r
boxplot(salary ~ rank, data = Salaries, main = "Salary by Rank", ylab = "Salary", xlab = "Rank",col="li
```



We can conduct an ANOVA to statistically test if there are differences in salaries based on ranks.

```
anova<-aov(salary ~ rank, data = Salaries)
summary(anova)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## rank          2 1.432e+11 7.162e+10   128.2 <2e-16 ***
## Residuals   394 2.201e+11 5.586e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As hypothesized, Professors, Associates, and Assistants have very different salaries. There is a statistically significant relationship between rank and salary, with a p-value less than 0.001. This implies that at least one rank has a salary that is significantly different from the others. Now, it's important to check the gender proportions for each rank

```
round(prop.table(table(Ass$sex))*100,3)
```

```
##
## Female   Male
## 16.418 83.582
```

```
round(prop.table(table(Assoc$sex))*100,3)
```

```
##
## Female   Male
## 15.625 84.375
```

```
round(prop.table(table(Prof$sex))*100,3)
```

```
##
## Female   Male
##  6.767 93.233
```

The proportion of females in the professor rank is much lower compared to the other two ranks. Therefore, analyzing the gender pay gap across the entire salary dataset is not the correct approach, as it would likely show a significant gap between men and women in favor of the former.

Therefore, we need to analyze the gender gap separately. The problem is that we have three datasets with a significant disparity between males and females, and these sets also have low sample sizes (particularly for Assistants and Associates). One possible approach to address this issue could be undersampling; however, given the small sample sizes, reducing the sample too much would lead to a considerable loss of information.

The idea is to execute both a t-test of Welch (because we don't assume the constraint that both groups have the same variance) and a bootstrap, for each rank.

-Assistants

```
M<-Ass %>%
  filter(sex=="Male")
F<-Ass %>%
  filter(sex=="Female")
t.test(M$salary,F$salary,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  M$salary and F$salary
## t = 1.0812, df = 12.941, p-value = 0.2993
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3258.378  9781.489
## sample estimates:
## mean of x mean of y
##  81311.46  78049.91
```

Since the p-value is greater than 0.05, we cannot reject the null hypothesis (H0). This means that there is no statistical evidence to suggest that the difference between male and female salaries is statistically significant.

```r
n_boot<-1000
bootstrap_diff<-numeric(n_boot)

for (i in 1:n_boot) {
  ms<-sample(M$salary,replace = TRUE)
  fs<-sample(F$salary,replace = TRUE)
  bootstrap_diff[i]<-mean(ms)-mean(fs)
}

ci<-quantile(bootstrap_diff, c(0.025, 0.975))

mean(bootstrap_diff)
```

```
## [1] 3207.595
```

```r
ci
```

```
##      2.5%     97.5%
## -2997.759  8572.524
```

Although the bootstrap difference in means indicates a difference of $3421 in favor of male salaries, the confidence interval includes zero. This suggests that we cannot confidently assert that there is a significant difference in salaries between the two groups. This aligns with the results of the previous t-test, which indicated that there was no statistical evidence of a significant difference in salaries.

-Associate

```r
M<-Assoc %>%
  filter(sex=="Male")
F<-Assoc %>%
  filter(sex=="Female")
t.test(M$salary,F$salary,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  M$salary and F$salary
## t = 1.0691, df = 10.781, p-value = 0.3084
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6762.142 19475.950
## sample estimates:
## mean of x mean of y
##   94869.7   88512.8
```

The p-value = 0.3084 indicates that there is insufficient evidence to reject the null hypothesis.

```r
n_boot<-1000
bootstrap_diff<-numeric(n_boot)

for (i in 1:n_boot) {
  ms<-sample(M$salary,replace = TRUE)
  fs<-sample(F$salary,replace = TRUE)
  bootstrap_diff[i]<-mean(ms)-mean(fs)
}

ci<-quantile(bootstrap_diff, c(0.025, 0.975))

mean(bootstrap_diff)
```

```
## [1] 6494.154
```

```r
ci
```

```
##      2.5%     97.5%
## -4878.208 17523.510
```

On average, male salaries are approximately $6436.87 higher than female salaries.However, the 95% confidence interval is [-4038.45, 17109.76]. This range includes zero, which suggests that we cannot conclude there is a statistically significant difference in salaries between the two groups.When considered alongside the earlier t-test results, which showed a p-value of 0.3084, both analyses indicate that while there is a positive difference in average salaries favoring males, this difference is not statistically significant.

-Prof

```r
M<-Prof %>%
  filter(sex=="Male")
F<-Prof %>%
  filter(sex=="Female")
t.test(M$salary,F$salary,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  M$salary and F$salary
## t = 1.0391, df = 22.451, p-value = 0.3098
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5119.769 15426.192
## sample estimates:
## mean of x mean of y
##  127120.8  121967.6
```

The p-value = 0.3098 indicates that there is insufficient evidence to reject the null hypothesis.

```
n_boot<-1000
bootstrap_diff<-numeric(n_boot)

for (i in 1:n_boot) {
  ms<-sample(M$salary,replace = TRUE)
  fs<-sample(F$salary,replace = TRUE)
  bootstrap_diff[i]<-mean(ms)-mean(fs)
}

ci<-quantile(bootstrap_diff, c(0.025, 0.975))

mean(bootstrap_diff)
```

```
## [1] 5091.874
```

```
ci
```

```
##      2.5%     97.5%
## -4132.349 14664.442
```

Male salaries are, on average, $5143.54 higher than female salaries. However, the 95% confidence interval is [-3973.191, 14371.651], which includes zero. This indicates that the difference is not statistically significant, meaning we cannot confidently conclude that there is a meaningful difference in salaries between men and women.

Now let's check if the t-test results changes if we consider the full Salaries dataset.

```
library(dplyr)
M<-Salaries %>%
  filter(sex=="Male")
F<-Salaries %>%
  filter(sex=="Female")
t.test(M$salary,F$salary)
```

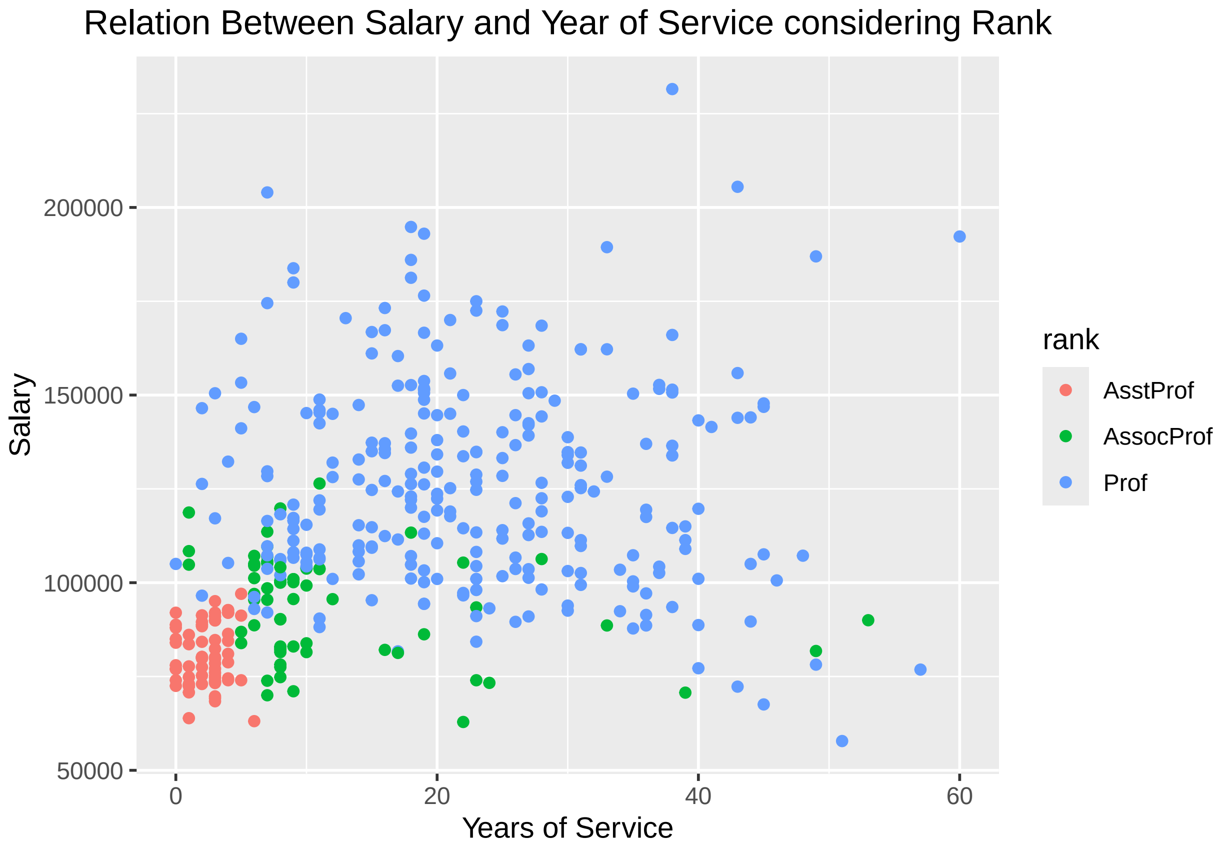
```
##
##  Welch Two Sample t-test
##
## data:  M$salary and F$salary
## t = 3.1615, df = 50.122, p-value = 0.002664
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5138.102 23037.916
## sample estimates:
## mean of x mean of y
##  115090.4  101002.4
```

The significant p-value (0.002664) and the confidence interval (which does not include zero) indicate that the difference in salaries is statistically significant. Therefore, we can conclude that there is evidence of a difference in the average salaries between the two groups.

In conclusion we can affirm that the existence of a statistical significant relation between salaries of male and females is strictly related to the distinction of the ranks.

- Relation between Salary and Years of Service.

```
library(ggplot2)
ggplot(Salaries,aes(x=yrs.service,y=salary,color=rank))+
  geom_point()+
  labs(x="Years of Service",y="Salary",title = "Relation Between Salary and Year of Service considering
  theme(plot.title = element_text(hjust = 0.5))
```



```
correlation <- cor(Salaries$yrs.service, Salaries$salary)
print(correlation)
```

```
## [1] 0.3347447
```

A correlation of 0.334 indicates a weak positive relationship between the two variables.

```
library(carData)
data("TitanicSurvival")
```

It's important to check that the einvoriment is correctly reading the data.

```
glimpse(TitanicSurvival)
```

```
## Rows: 1,309
## Columns: 4
## $ survived       <fct> yes, yes, no, no, no, yes, yes, no, yes, no, no, yes, y~
## $ sex            <fct> female, male, female, male, female, male, female, male,~
## $ age            <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, 63.~
## $ passengerClass <fct> 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, 1st, ~
```

Every variable of the dataset is correctly expressed. Now let's check if there are any missing values.

```
sum(is.na(TitanicSurvival))
```

```
## [1] 263
```

Titanic Survival datasets presents 263 missing values divide in such way:

```
colSums(is.na(TitanicSurvival))
```

```
##       survived            sex            age passengerClass 
##              0              0            263              0
```

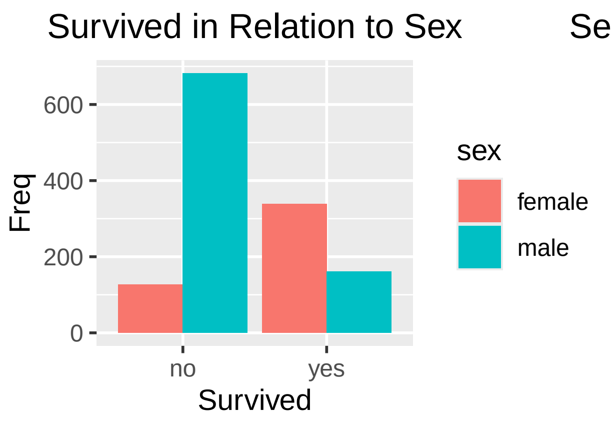
Excluding now all passengers with missing values in the age variable wouldn't be the right choice. In fact, we might need those records.

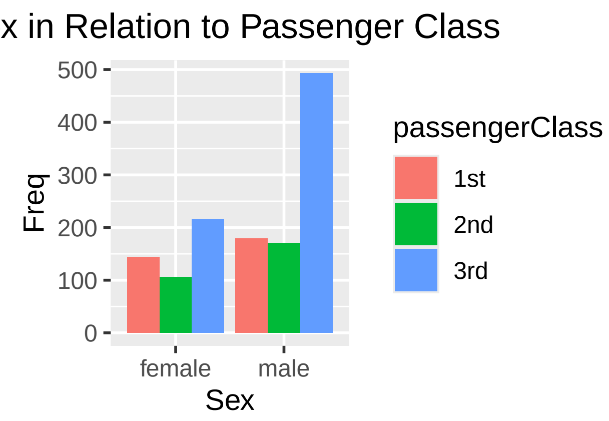
By using the summary function, we can have a general idea of the characteristics of the data. The result varies in respect to the type of variable, such as double, integer, factor etc.

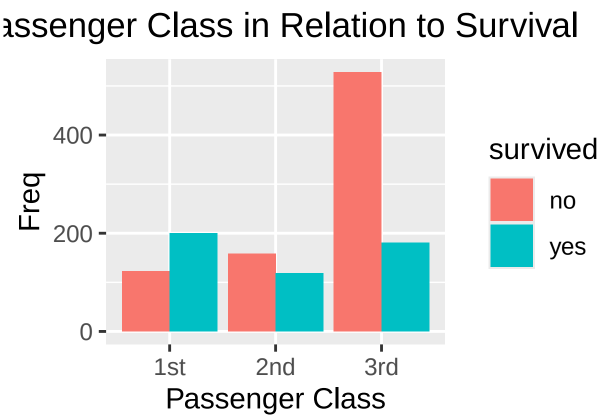
```
summary(TitanicSurvival)
```

```
##  survived      sex           age          passengerClass
##  no :809   female:466   Min.   : 0.1667   1st:323       
##  yes:500   male  :843   1st Qu.:21.0000   2nd:277       
##                         Median :28.0000   3rd:709       
##                         Mean   :29.8811                 
##                         3rd Qu.:39.0000                 
##                         Max.   :80.0000                 
##                         NA's   :263
```

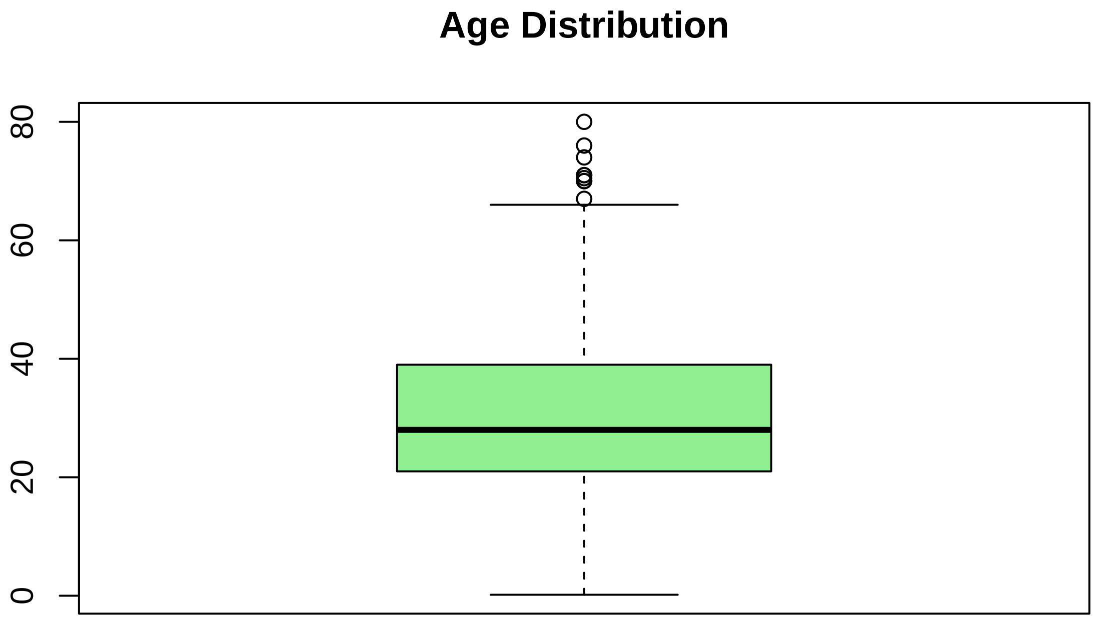
```
p1<-ggplot(TitanicSurvival,aes(x=survived,fill=sex))+
  labs(x="Survived",y="Freq",title = "Survived in Relation to Sex")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")
p2<-ggplot(TitanicSurvival,aes(x=sex,fill=passengerClass))+
  labs(x="Sex",y="Freq",title = "Sex in Relation to Passenger Class")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")
p3<-ggplot(TitanicSurvival,aes(x=passengerClass,fill=survived))+
  labs(x="Passenger Class",y="Freq",title = "Passenger Class in Relation to Survival")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_bar(position = "dodge")

grid.arrange(p1,p2,p3,nrow=2,ncol=2)
```



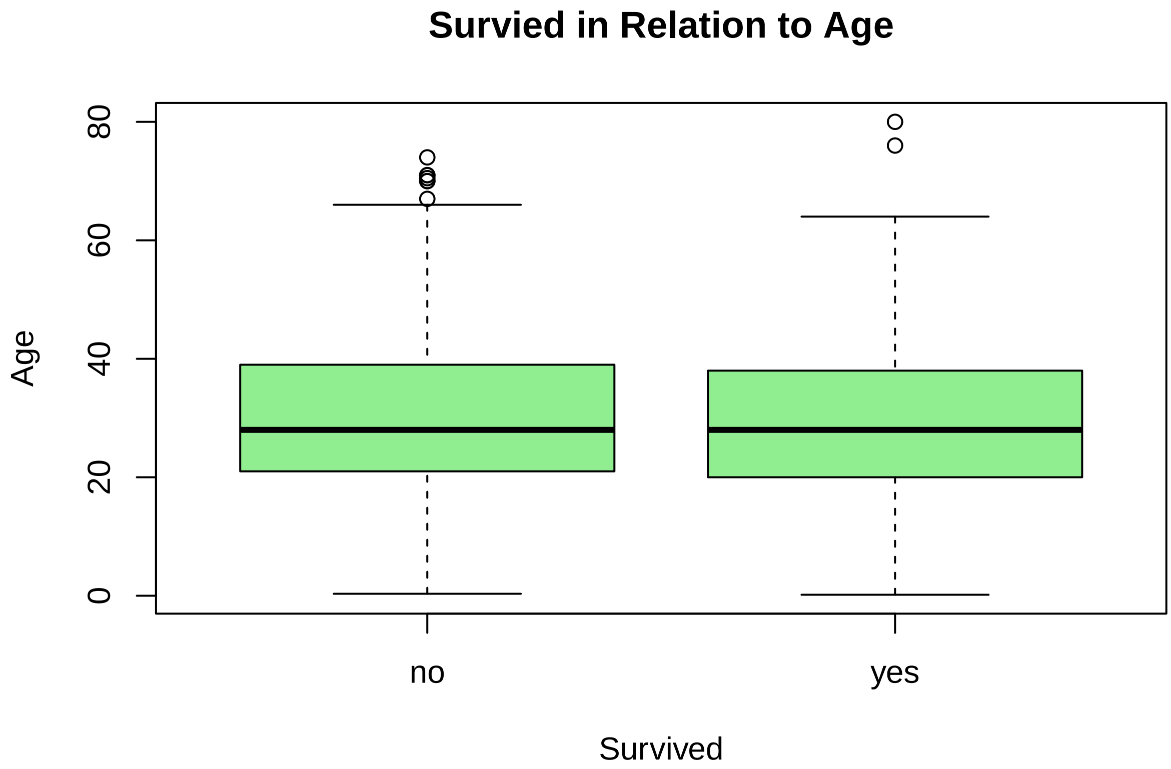




```
boxplot(TitanicSurvival$age,col="lightgreen",main="Age Distribution")
```



```
boxplot(age ~survived, data=TitanicSurvival,xlab="Survived",ylab="Age",main="Survied in Relation to Age
```



From the summary, we see that a majority of the passengers (809) did not survive, while 500 survived. The dataset is also imbalanced in terms of gender, with 843 males compared to 466 females. Ages range from a few months (0.17 years) to 80 years, with an average age around 30. The bar charts reveal some notable patterns. The Survival by Sex chart shows that a significantly higher proportion of males did not survive, while females had a much better survival rate. This suggests that gender played a significant role in survival, possibly reflecting a "women and children first" protocol during the disaster. The Sex in Relation to Passenger Class chart shows that males were predominantly in 3rd class, while females were more evenly distributed across all classes. This imbalance might explain, in part, why survival rates differ by gender, as passenger class itself had a strong association with survival. The Passenger Class in Relation to Survival chart emphasizes that survival rates varied significantly across classes. First-class passengers had a noticeably higher survival rate compared to those in third class, where the majority of passengers did not survive. Finally, the Age by Survival box plot reveals that age did not drastically differ between survivors and non-survivors, as the median ages are similar across both groups. Although there are a few older outliers among non-survivors, the box plot suggests that age may not have been a decisive factor in survival compared to gender and class.

To check if there is independence between Survived and Class, we can use a chi-square test.

```
c<-chisq.test(table(TitanicSurvival$survived,TitanicSurvival$passengerClass))
c
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(TitanicSurvival$survived, TitanicSurvival$passengerClass)
## X-squared = 127.86, df = 2, p-value < 2.2e-16
```

The Chi-Square statistic (127.86) indicates that there is a large difference between observed and expected counts, suggesting a strong association between two survived and passenger class. The p-value is extremely small. This leads us to reject the null hypothesis of independence between survived and passenger class.

```
c$expected
```

```
##
##            1st     2nd      3rd
##   no  199.6234 171.194 438.1826
##   yes 123.3766 105.806 270.8174
```

These are the theoretical counts under the assumption of independence

```
c$observed
```

```
##
##       1st 2nd 3rd
##   no  123 158 528
##   yes 200 119 181
```

Meanwhile these are the real observed data. There is a Low similarity between these two tables, suggesting a significant association, indicating they are likely not independent.

```
c$stdres
```

```
##
##              1st        2nd        3rd
##   no  -10.110480  -1.837589  10.254442
##   yes  10.110480   1.837589 -10.254442
```

The standardized residuals tell how far each observed value is from the expected.Positive residuals suggest that the values is higher than expected, the opposite happens for the negative residuals. An absolute value >= 2 indicates significant deviation from independence for that cell (in these case all 1st and 3st class passengers with yes and non combination of survival). When the abs values is smaller than 2 observed and expected values are close, suggesting independence for that cell (2nd class passengers).