Tutorial Business Analytics

R Tutorial 1 - Solution

Exercise 1.1 Loading a data set and statistics

Note: Please load the library using *library(tidyverse)* command.

a) Read the CSV file "LaborSupply1988.csv" into a tibble `df`.

```
df = read_csv("PathToFile//LaborSupply1988.csv")
```

b) How many attributes (columns) and observations (rows) does `df` have?

```
The tidyverse way
glimpse(df)
The other way
str(df)

nrow(df)
ncol(df)
```

c) Which attributes does the data set have?

```
names(df)
# lnhr: log of annual hours worked
# lnwg: log of hourly wage
# kids: number of children
# age: age
# disab: bad health
```

d) List the first rows of the data set.

```
head(df, n=20)
```

e) What is the value range of the attribute - `age`?

```
The tidyverse way
summarise(df, min_age=min(age), max_age=max(age))
The other way
summary(df$age)
min(df$age)
max(df$age)
range(df$age)
```

f) Calculate the average of annual hours worked by the labourers with 0, 1, 2, ... 6 kids each.

```
The tidyverse way
df %>% group_by(kids) %>% summarise(mean_lnhr=mean(lnhr))
The other way
mean(df[df$kids == 0,]$lnhr)            # repeat with 1,2,...,6
```

g) Calculate the average number of `kids` of the 40 year old.

```
The tidyverse way
df %>% filter(age == 40) %>% summarise(mean_kids=mean(kids))
The other way
mean(df[df$age == 40, ]$kids)
```

Exercise 1.2 Plotting

a) Plot a histogram of the attribute `age`. What is the most frequent age?

```
hist(df$age)
df %>% group_by(age) %>% summarise(count=n()) %>% arrange(desc(count))
```

The most frequent age is 39.

b) Plot the average number of `kids` against the `age` and interpret the resulting graph. Underpin your observation using a statistical method.

```
The tidyverse way
plot(df %>% group_by(age) %>% summarise(avg_kids=mean(kids)))
The other way
plot(aggregate(x=df$kids, by=list(df$age), FUN=mean))
```

The average number of kids decreases with increasing age.

```
cor(df$kids, df$age)
```

The two attributes are correlated negatively.

c) Plot the log of hourly wage (`lnwg`) against the `age`.

```
plot(df$age, df$lnwg)
```

d) Plot the mean of the log of hourly wage (`lnwg`) against the `age`. How are they correlated? Also compute the correlation.

```
The tidyverse way
plot(df %>% group_by(age) %>% summarise(avg_lnwg=mean(lnwg)))
The other way
plot(aggregate(x=df$lnwg, by=list(df$age), FUN=mean))
cor(df$lnwg, df$age)
```

e) Plot `lnhr` against the `age` with different colors for `disab=0` and `disab=1`.

```
plot(df$age, df$lnhr, pch=df$disab+1, col=c("red", "blue")[df$disab+1])
```

f) Plot a boxplot of the log of annual hours worked (`lnhr`) against the number of `kids`. What could be observed regarding mean and variance? Is the observation meaningful for large values of `kids`?

```
boxplot(df$lnhr ~ df$kids)
```

The mean increases with an increasing number of kids, while the variance decreases.

```
hist(df$kids, breaks=(max(df$kids)-min(df$kids)))
```

For values of 5 and 6, only two observations exist. Hence the observation is not very meaningful.