Tutorial Business Analytics

R Tutorial 1 - Solution

Exercise 1.1 Loading and describing a data set

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)
```

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

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

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

Tutorial Business Analytics

Homework 1 - Solution

Exercise 2.1: Describing the beer consumption on the Oktoberfest

a) Read the provides CSV file (“Oktoberfest.csv”) and store it in a tibble named *oct.*

```
oct = read_csv("Oktoberfest.csv")
```

b) Which attributes does the data set have?

```
names(oct)
```

c) What was the price of a beer in 1995?

```
Base R Solution
oct[oct$Year == 1995,]$Beer_Price

TidyVerse Solution
oct %>% filter(Year == 1995) %>% select(Beer_Price)
```

d) Based on the data set, when did the city of Munich first recorded the beer price?

```
min(oct$Year)
```

e) What is the value range of the attribute – *Visitors_Total* describing the total number of visitors in million in the corresponding year?

```
TidyVerse Solution
summarize(oct, min_vis=min(Visitors_Total), max_vis=max(Visitors_Total))

Base R Soluation
min(oct$Visitors_Total)
max(oct$Visitors_Total)
range(oct$Visitors_Total)
```

f) Plot and describe the beer consumption over the years

```
Base R Solution
plot(oct$Year, oct$Beer_Consumption, type=’line’)

ggplot2 Solution
ggplot(oct, aes(x=Year, y=Beer_Consumption)) + geom_line()
```

The plots indicates that the beer consumption increased over the years.

g) The number of visitors could provide an explaination to this observation. Create a scatter-plot that shows the number of visitors per year. Subsequently, calculate a statistic to validate or reject this explanation.

```
Base R Solution
plot(oct$Year, oct$Visitors_Total)

ggplot2 Solution
ggplot(oct, aes(x=Year, y=Visitors_Total)) + geom_point()

cor(oct$Visitors_Total, oct$Beer_Consumption)
```

The plots show that the number of visitors varies between 5.5 and 7 million people. Moreover, it indicates that, on average, the number decreases. Due to the opposing trends, it cannot be an explanation. The negative correlation coefficients supports this.
Caution: This interpretation and the overall approach is not meaningful from a statistical persepective.

Exercise 2.2: Describing the beer price on the Oktoberfest

The goal of this exercise is to use *dplyr* for summarizing the data set.

a) What was the average beer price from 2000 to 2007?

```
oct %>% filter(Year >= 2000, Year <= 2007) %>% summarize(avg_prize =
mean(Beer_Price))
```

b) What was the variance of the beer price within this time frame?

```
oct %>% filter(Year >= 2000, Year <= 2007) %>% summarize(var_price =
var(Beer_Price))
```

c) Add a new variable *difference* using the *mutate* function that describes the difference between the beer price of a year and the previous year.

```
oct = oct %>% mutate(difference = Beer_Price - lag(Beer_Price))
```

d) Plot these differences per year using ggplot2.

```
ggplot(tail(oct, -1), aes(x = Year, y = difference)) + geom_line()
```