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()
```