Exercise 3: House Prices

In this exercise, you will work with the house prices dataset obtained from Kaggle. The dataset contains 81 columns describing (almost) every aspect of residential homes in Ames, Iowa.

```
library("readr")
library("dplyr")
library("ggplot2")
```

3.a Save the data file in `data` folder of your project. Load the dataset into the global environment by using function `read_csv` and assign it to variable `house_prices`. Make sure that the class of `house_prices` is a tibble.

```
house_prices<-read_csv("data/house_prices.csv")
class(house_prices)
```

3.b You will work with only four variables, namely, `LotArea`, `KitchenQual`, `LotShape`, and `SalePrice`, which indicate area (in square feet), kitchen quality, general shape, and sale price (in dollars) of property. Modify `house_prices` to have only these four columns, as well as transform `LotArea` from square feet into square meters. Bonus: try to use only one `dplyr` function.

```
house_prices<-house_prices%>%
  transmute(LotArea_m=(LotArea/10.76391), KitchenQual, LotShape, SalePrice)
```

3.c For how many lots the sale price was greater than its mean value?

```
house_prices%>%
  filter(SalePrice>mean(SalePrice))%>%
  summarise(n_lots_high_price=n())
```

The sale price of 560 lots were greater than its mean value.

3.d Display the average sale price for each kitchen quality level.

Note, levels of `KitchenQual` correspond to the followin values:

```
- Ex -- Excellent,
- Gd -- Good,
- TA -- Typical/Average,
- Fa -- Fair
- Po -- Poor

house_prices%>%
  group_by(KitchenQual)%>%
  summarise(Average_sale_price= mean(SalePrice))
```

3.e Display ten randomly selected observations from the dataset. What happens if you execute your code a few times? How can you make sure that each execution returns the exact 10 rows?

```
house_prices%>%
  sample_n(size=10)
```

If the code is executed a few times can be noticed that the rows selected change.

```
set.seed(32)
house_prices%>%
  sample_n(size=10)
```

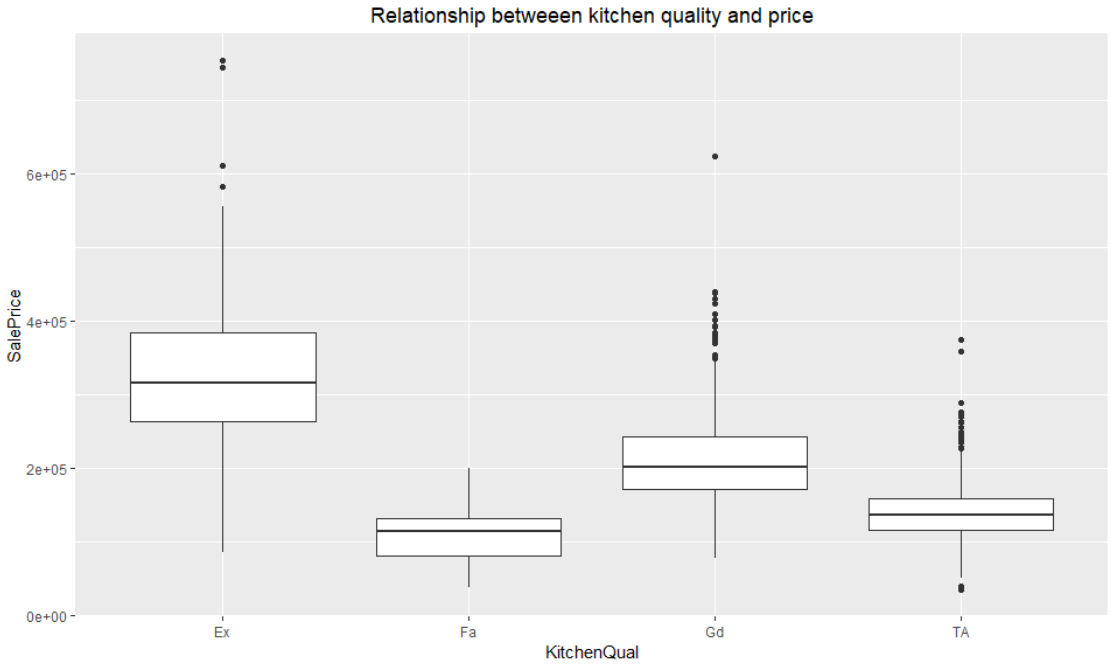
To be able to reproduce the same results it is possible to set a seed. The number of seed specify a vector of random numbers that is going to be used.

3.f Print out the minimum price of observations for which shape is regular (i.e., "Reg"), and the kitchen has excellent quality.

```
house_prices%>%
  filter(LotShape=="Reg" & KitchenQual=="Ex")%>%
  summarise(Min_price=min(SalePrice))
```

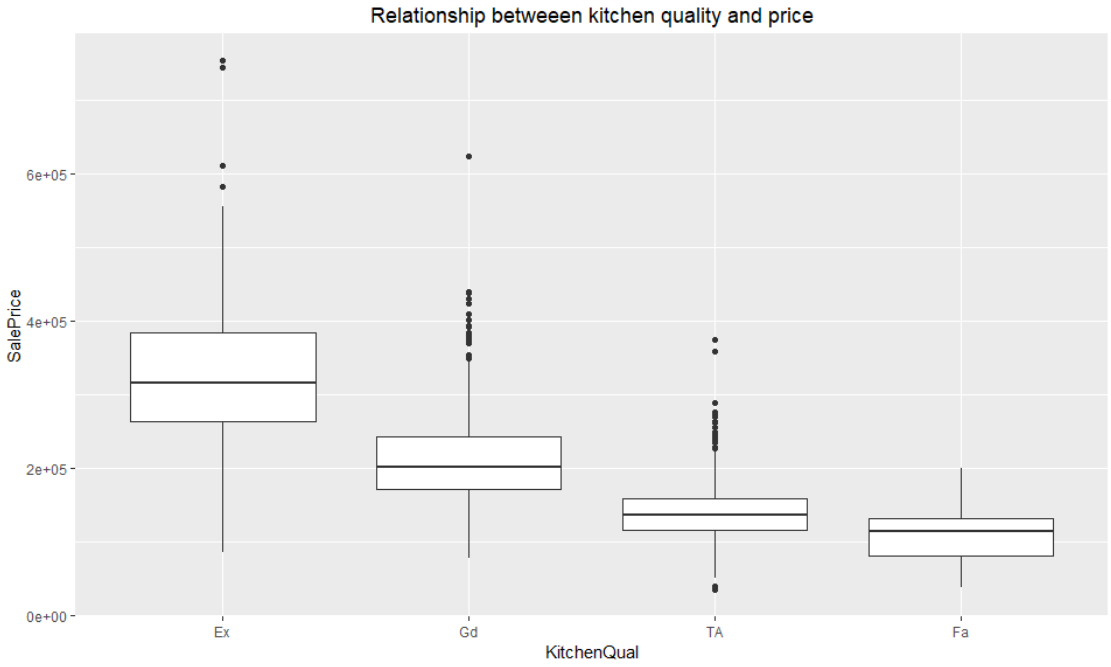
3.g Produce boxplots with kitchen quality as the x-axis and the price as the y-axis. Regroup kitchen quality in the following order: excellent, good, typical/average, and fair. What can you say about the relationship between kitchen quality and the price?

```
house_prices%>%
  ggplot(mapping=aes(x=KitchenQual, y=SalePrice))+
  ggtitle("Relationship betweeen kitchen quality and price")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_boxplot()
```




```
house_prices%>%
  ggplot(mapping=aes(x=KitchenQual, y=SalePrice)) +
```

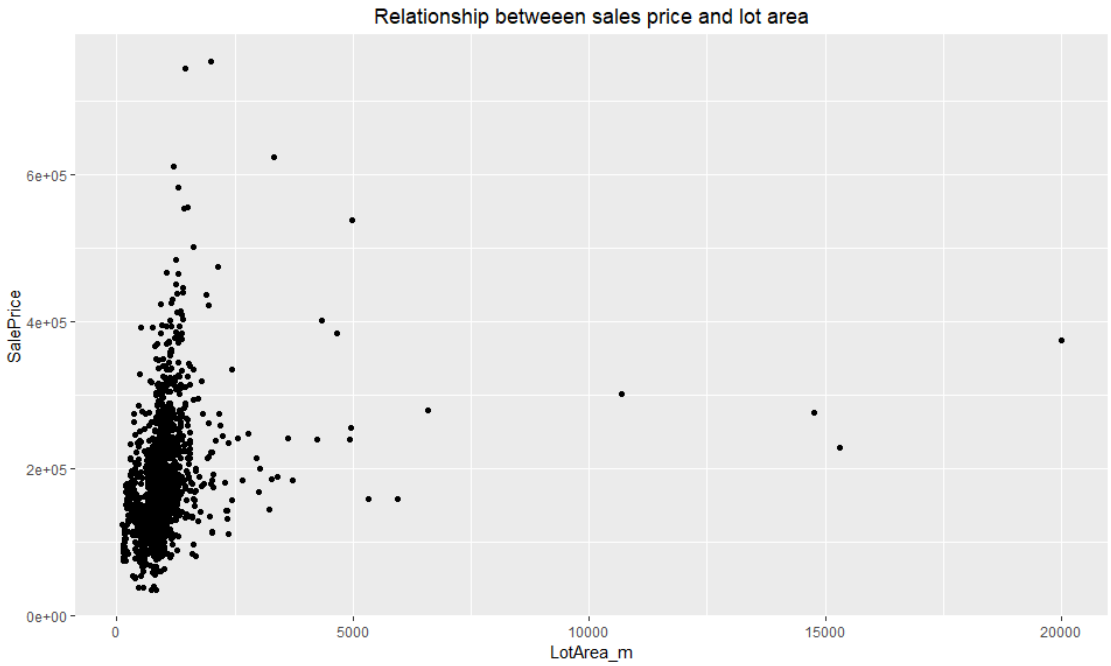
```
  geom_boxplot(mapping=aes(x=factor(KitchenQual, level= c("Ex", "Gd", "TA",
"Fa"))))+
  ggtitle("Relationship betweeen kitchen quality and price")+
  theme(plot.title = element_text(hjust = 0.5))
```



The price of a residential home in Ames, Iowa decreases whith the level quality of the kitchen.

3.h Draw a scatter chart to investigate the dependence between `LotArea` and `SalePrice`. Further, use different colors depending on the kitchen quality, and different shapes depending on the shape of the property.

```
house_prices%>%
  ggplot(mapping = aes(x=LotArea_m, y=SalePrice))+
  ggtitle("Relationship betweeen sales price and lot area")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_point()
```




```
house_prices%>%
  ggplot(mapping = aes(x=LotArea_m, y=SalePrice, color=KitchenQual,
shape=LotShape))+
  ggtitle("Relationship betweeen sales price and lot area")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_point()+
  scale_shape_manual(values=c(0, 3, 16,17))+
  scale_color_manual(values=c("red", "green", "orange","bisque3"))
```