Homework 1

MSSC 6010- Computational Probability

Henri Medeiros Dos Reis

September 6, 2023

Question 1. (1.21.5) From book. Vectors, sequences, and logical operators.

- *(a)* Assign the names x and y to the values 5 and 7, respectively. Find x^y and assign the result to z. What is the valued stored in z?

```
x <- 5
y <- 7
(z <- x**y)
```

```
[1] 78125
```

The value of z is 78125

- *(b)* Create the vectors $u = (1, 2, 5, 4)$ and $v = (2, 2, 1, 1)$ using the `c()` function

```
u <- c(1,2,5,4)
v <- c(2,2,1,1)
```

- *(c)* Provide R code to find which component of u is equal to 5.

```
which(u==5)
```

```
[1] 3
```

- *(d)* Provide R code to give the components of v greater than or equal to 2.

```
which(v>=2)
```

```
[1] 1 2
```

- *(e)* Find the product $u \times v$. How does R perform the operation?

```
u*v
```

```
[1] 2 4 5 4
```

This operation is performed element wise.

- *(f)* Explain what R does when two vectors of unequal length are multiplied together. Specifically, what is `u` $\times$ `c(u, v)`?

```
u*c(u,v)
```

```
[1]  1  4 25 16  2  4  5  4
```

The value is the shorter vector get multiplied again to match the elements that are left. Once all the elements in the first vector are over, then it starts again.

- *(g)* Provide R code to define a sequence from 1 to 10 called G and subsequently to select the first three components of G

```
G <- seq(1,10)
G[1:3]
```

```
[1] 1 2 3
```

- *(h)* Use R to define a sequence from 1 to 30 named J with an increment of 2 and subsequently to choose the first, third, and eighth values of J.

```
J <- seq(1,30,2)
J[c(1,3,8)]
```

```
[1]  1  5 15
```

- *(i)* Calculate the scalar product (dot product) of $q = (3, 0, 1, 6)$ by $r = (1, 0, 2, 4)$.

```
q <- c(3,0,1,6)
r <- c(1,0,2,4)
(q%*%r)
```

```
     [,1]
[1,]   29
```

- *(j)* Define the matrix X whose rows are the u and v vectors from part (b).

```
(X <- rbind(u,v))
```

```
  [,1] [,2] [,3] [,4]
u    1    2    5    4
v    2    2    1    1
```

- *(k)* Define the matrix Y whose columns are the u and v vectors from part (b).

```
(Y <- cbind(u,v))
```

```
     u v
[1,] 1 2
[2,] 2 2
[3,] 5 1
[4,] 4 1
```

- *(l)* Find the matrix product of X by Y and name it W.

```
(W <- X%*%Y)
```

```
   u  v
u 46 15
v 15 10
```

- *(m)* Provide R code that computes the inverse matrix of W and the transpose of that inverse.

```
(invW <- solve(W))
```

```
            u           v
u  0.04255319 -0.06382979
v -0.06382979  0.19574468
```

```
(invWT <- t(invW))
```

```
            u           v
u  0.04255319 -0.06382979
v -0.06382979  0.19574468
```

Question 2. (2.10.5) From book. The data frame `VIT2005` in the `PASWR2` package contains descriptive information and the appraised total price (in euros) for apartments in Vitoria, Spain.

```
library(MASS)
library(PASWR2)
library(dplyr)
```

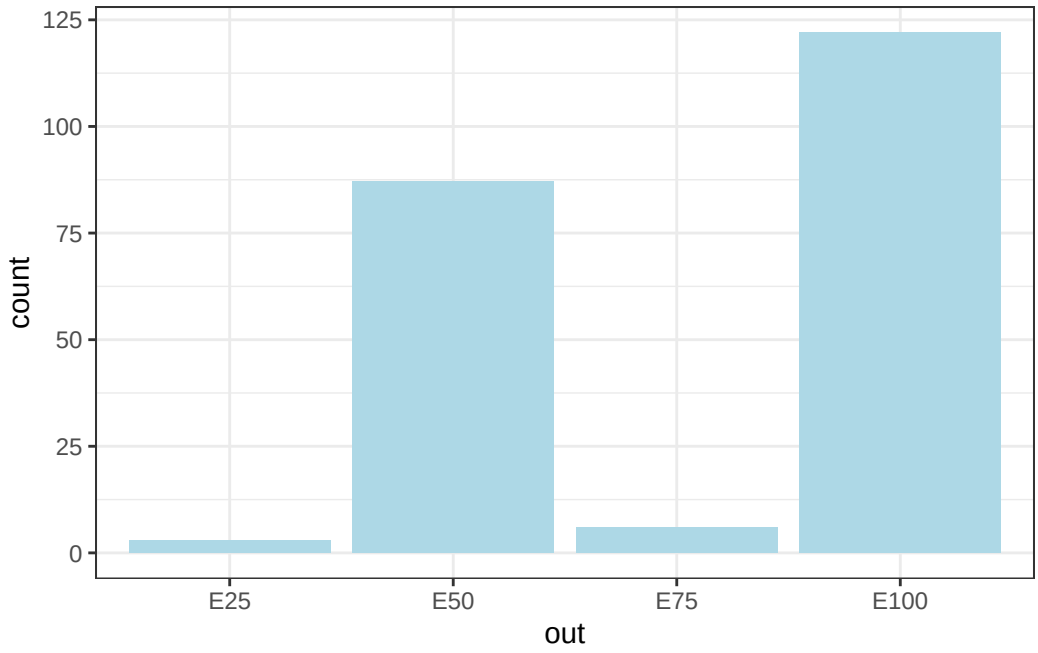
- *(a)* Create a frequency table, a piechart, and a barplot showing the number of apartments grouped by the variable `out`. For you, which method conveys the information best?

```
head(VIT2005,5)
summary(VIT2005)
```

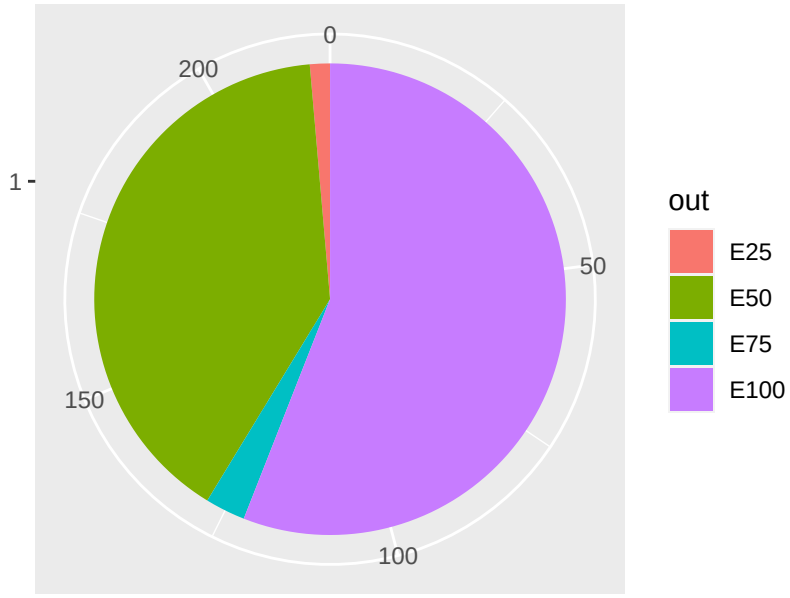
```
VIT2005$out <- factor(VIT2005$out, levels = c("E25", "E50", "E75", "E100"))
summary(VIT2005$out)
```

```
 E25  E50  E75 E100
   3   87    6  122
```

```
ggplot(data = VIT2005, aes(x = out)) +
  geom_bar(fill="lightblue") +
  theme_bw()
```



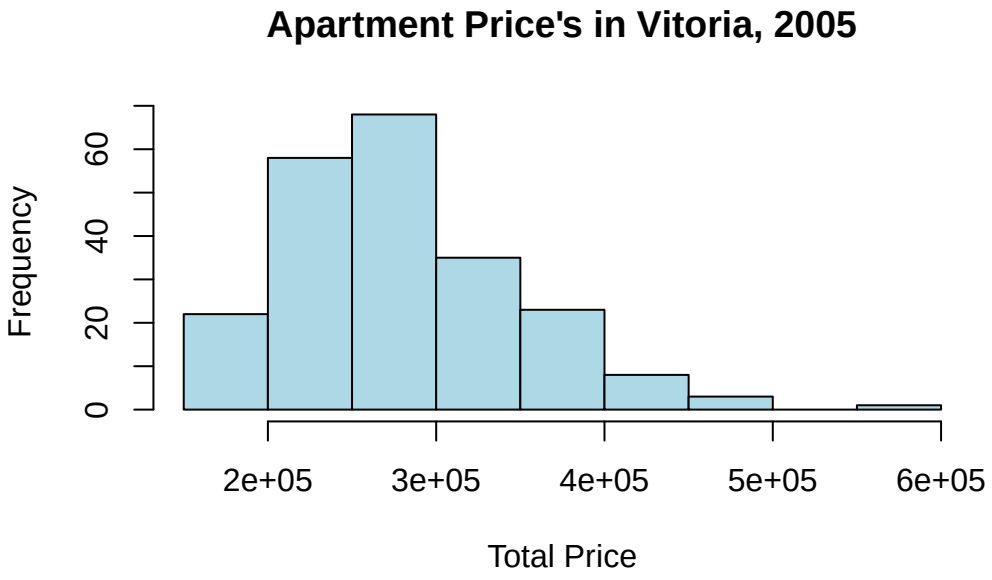

```
ggplot(data = VIT2005, aes(x = factor(1), fill = out)) +
  geom_bar()+
  coord_polar(theta = "y") +
  labs(x = "",y = "")
```



The method that conveys the information the best is the bar plot.

- *(b)* Characterize the distribution of the variable `totalprice`.

```
hist(VIT2005$totalprice, col = "lightblue", xlab = "Total Price",
     main = "Apartment Price's in Vitoria, 2005")
```



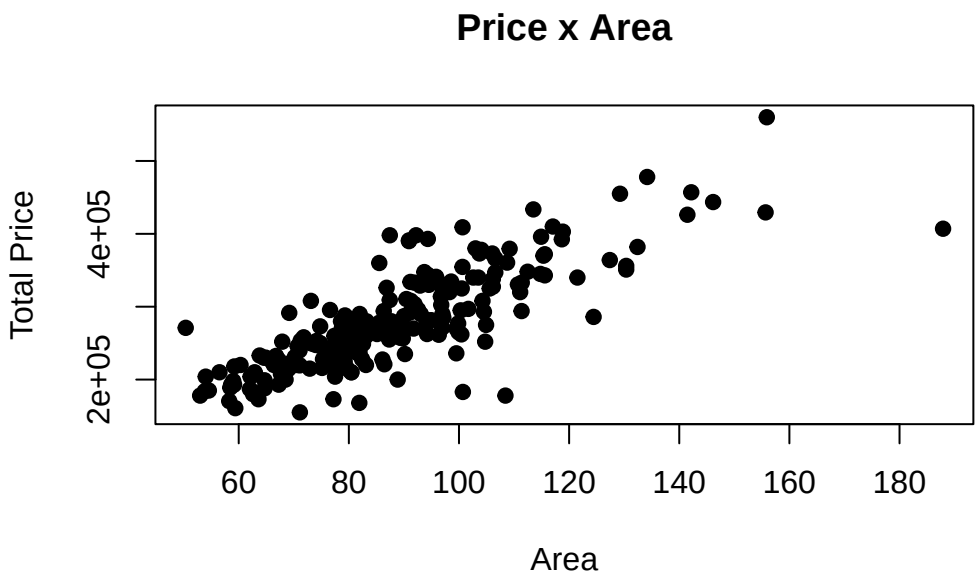

```
summary(VIT2005$totalprice)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 155000  228500  269750  280742  328625  560000
```

As it is possible to see, this distribution is skewed to the right. It has 1 outlier, of the value 560000$, median of 269750$, and mean 280742$

- *(c)* Characterize the relationship between `totalprice` and `area`.

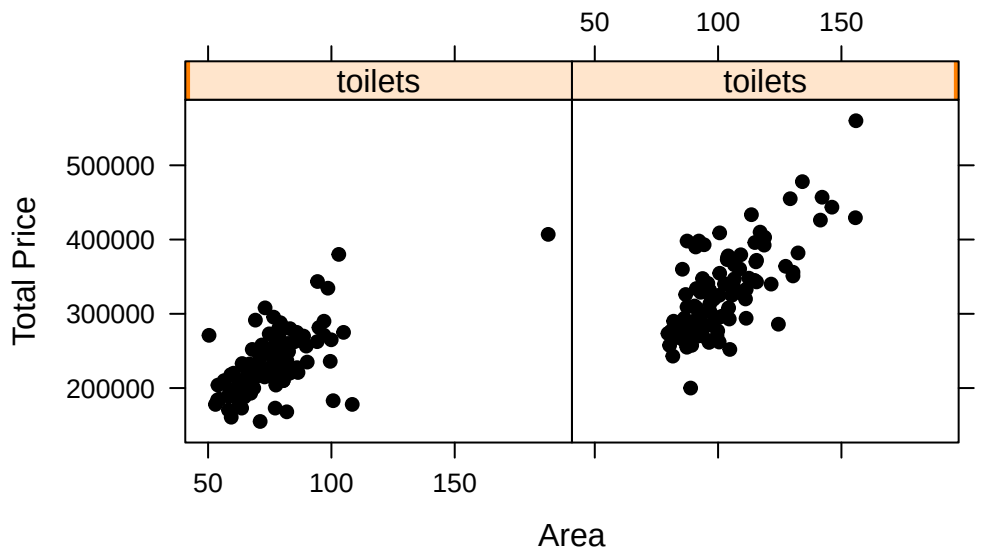
```
# Checking for a rtlationship between totalprice and area
plot(VIT2005$area, VIT2005$totalprice, pch = 19, xlab = "Area",
     ylab = "Total Price", main = "Price x Area")
```



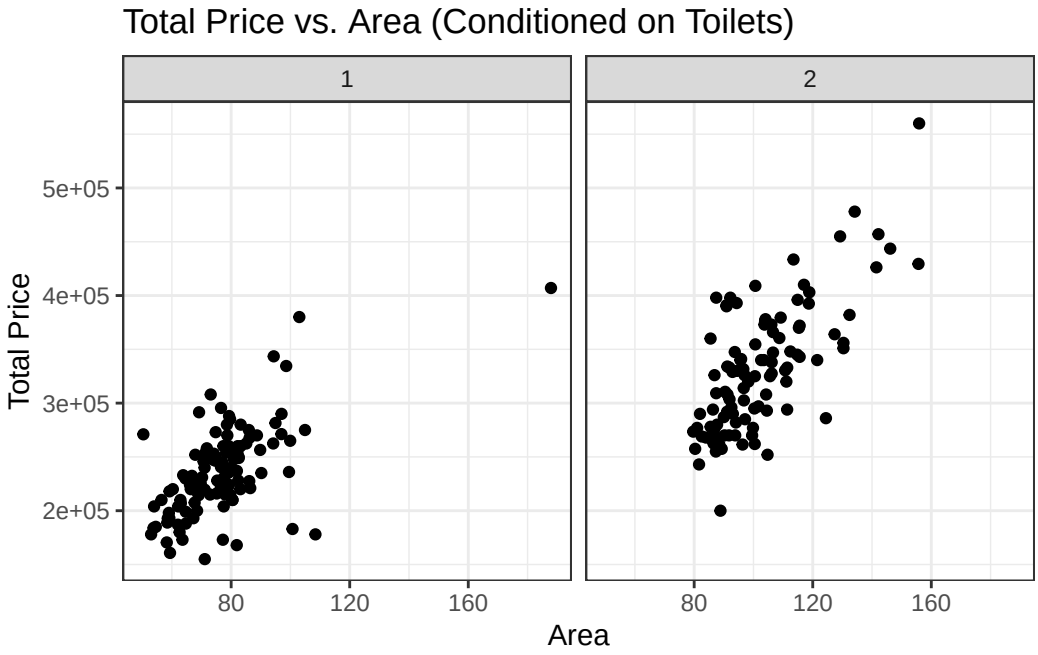
There exists a positive linear relationship between Area and Price.

- *(d)* Create a Trellis plot of `totalprice` versus `area` conditioning on `toilets`. Create the same graph with `ggplot2` graphics. Are there any outliers? Ignoring any outliers, between what two values of `area` do apartments have both one and two bathrooms?

```
# Trellis plot
xyplot(totalprice ~ area | toilets, data = VIT2005, layout = c(2, 1),
       main = "Total Price vs. Area (Conditioned on Toilets)",
       xlab = "Area", ylab = "Total Price",
       par.settings = simpleTheme(pch = 19, col = "black"))
```




```
# Same plot using ggplot
ggplot(data = VIT2005, aes(x = area, y = totalprice)) +
  geom_point() +
  labs(x = "Area", y = "Total Price",
       title = "Total Price vs. Area (Conditioned on Toilets)") +
  theme_bw() +
  facet_grid(~toilets)
```



Let's check for outliers

```
# Calculate outliers using IQR method
toilets_outliers <- VIT2005 %>%
  group_by(toilets) %>%
  summarize(
    lower_fence = quantile(totalprice, 0.25) - 1.5 * IQR(totalprice),
    upper_fence = quantile(totalprice, 0.75) + 1.5 * IQR(totalprice)
  )
outliers <- VIT2005 %>%
  left_join(toilets_outliers, by = "toilets") %>%
  filter(totalprice < lower_fence | totalprice > upper_fence)
outliers$toilets
```

```
[1] 2 1 1 1 1
```

As we can see, there are five outliers, four with 1 toilet and one with 2 toilets.

Now, let's see in what range of price the apartments can have either 1 or 2 toilets.

```
outliers <- outliers[-c(16,17)]

filtered <- setdiff(VIT2005,outliers)
one_t <- subset(filtered, subset = toilets==1)
two_t <- subset(filtered, subset = toilets==2)

max(one_t$area)
```

```
[1] 108.44
```

```
min(two_t$area)
```

```
[1] 79.68
```

The overlapping range is from 80 squared meters to 110 squared meters.

- *(e)* Use the `area` values reported in (d) to create a subset of apartments that have both one and two bathrooms. By how much does an additional bathroom increase the appraised value of an apartment? Would you be willing to pay for an additional bathroom if you lived in Vitoria, Spain?

```
both_t <- subset(filtered, subset = area >= 80 & area <= 110)
my_diff <- tapply(both_t$totalprice, both_t$toilets, mean)
my_diff[2]-my_diff[1]
```

```
       2
67606.76
```

An additional bathroom averages to be 67,607 increase in the price. I would be willing to pay extra in order to have an additional bathroom.