Exam, R lectures,; Prof. An Carbonez;
Be sure to regularly save your work. This is your own responsibility!

At the end of the exam:

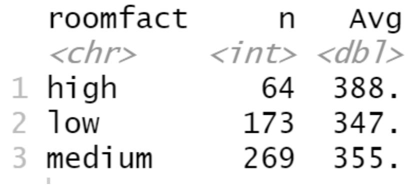
(i) Write your name on this document and hand in.
(ii) BE SURE THAT THE FIRST LINE OF YOUR R SCRIPT CONTAINS YOUR NAME
(iii) Upload the R script file (with extension .R) on Toledo > Assignment > upload R part.

Part 1: R part (total points: 10))

Use the *dataset boston2.xlsx.* This data set contains information on crime rates in the districts of Boston. Next variables are available:

Name variabele	**Description**
obs	Observation number
crim	Crime rate per district
chas	Variable which indicates whether the Charles River passes through the district ✓ Chas=1 : river passes the district ✓ Chas=0: river does not pass the district
rm	Average number of rooms in a house
roomfact	Categorical variable indicating the type of house. ✓ low : average number of rooms < 6 ✓ medium : average number of rooms =6 ✓ high : average number of rooms > = 7
dis	Average distance to the industrial area of Boston
rad	Average distance to the nearest highway
zn	Proportion of residential housing
indus	Proportion of industry in that district
age	Average age (in years) of the people living in that district
crimcat	✓ 1: if the crime rate is high ✓ 0: if the crime rate is low
international	Variable indicating the type of district (this is a continuous variable). It describes the proportion of international residents in this district.
internationfact	Categorical variable indicating the type of district. This variable describes whether there are a lot of international residents in this districts. ✓ low : ✓ medium : ✓ high : ✓ very_high

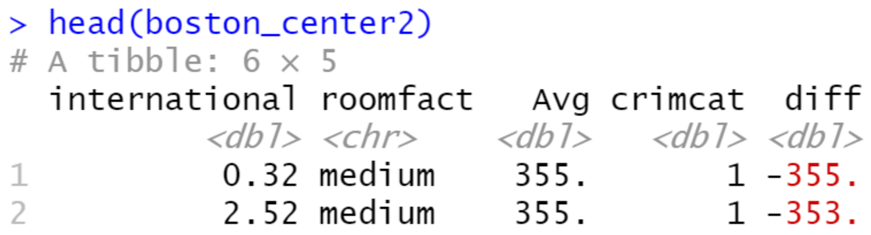
#1. Compute averages of international (proportion of international residents) by type of house (= variable roomfact) and corresponding number of districts. You have to use dplyr package. Save the output in summary1;



```
  roomfact      n   Avg
  <chr>     <int> <dbl>
1 high         64  388.
2 low         173  347.
3 medium      269  355.
```

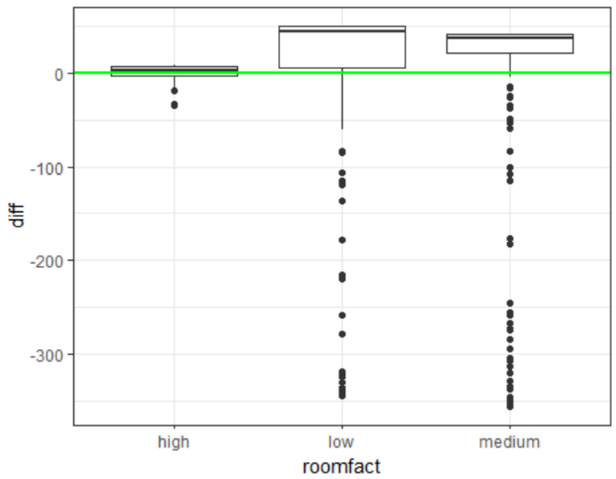
# 2. Merge summary1 with the original data set boston2 in a logical way. We want to center the values of international in the following way.

- Compute difference between the international value in a district and its average
- We select only the variables roomfact, average, international, diff, crimcat
- Sort this dataframe by diff (small to large)
- This is the dataframe boston_center2



```
> head(boston_center2)
# A tibble: 6 × 5
  international roomfact   Avg crimcat  diff
          <dbl> <chr>    <dbl>   <dbl> <dbl>
1          0.32 medium    355.       1 -355.
2          2.52 medium    355.       1 -353.
```

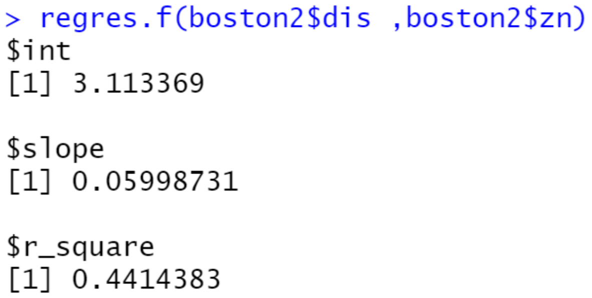
# 3. make a box plot of the differences (use ggplot2)
# add a horizontal green reference line (at a diff value of 0) (hint: look at the help of geom_abline)



# 4. Create a function **regres.f** (with as input parameter *resp* (= the response variable) and *expl* (= the explanatory variable X)) for performing linear regression

# This function gives as output : the intercept, slope + R² value (*hint: look at summary.lm*)

Apply this function to investigate the linear regression between response variable ‘distance to industrial area ‘(dis) and the explanatory variable ‘proportion of residential housing’ (zn).



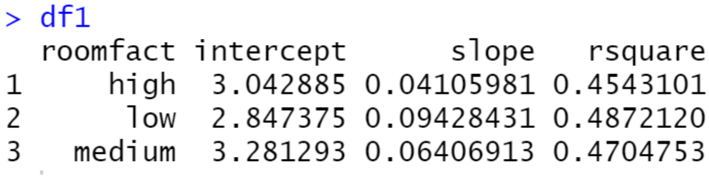
```
> regres.f(boston2$dis ,boston2$zn)
$int
[1] 3.113369

$slope
[1] 0.05998731

$r_square
[1] 0.4414383
```

5. # For answering this question, please make use of the function **regres.f** which was created in question 4. Use regres.f for every level of roomfact. (use also distance to industrial area (dis) as response variable and proportion of residential housing as explanatory variable. (zn).

Do this in such a way that you obtain a data frame df1 with the following information.



```
> df1
  roomfact intercept      slope   rsquare
1     high  3.042885 0.04105981 0.4543101
2      low  2.847375 0.09428431 0.4872120
3   medium  3.281293 0.06406913 0.4704753
```