**MASTER IN DATA SCIENCE (FIB-UPC).**

**ACADEMIC YEAR 22-23 Q1** – **PARTIAL EXAM**

**Statistical Inference and Modelling (SIM) .**

**Date: 3/Nov/2022 18:30-20:30 h Classrooms C6S302 - Group 11, C6S303-Group 12**

**Professor**: Lídia Montero and Josep Franquet

**Rules for quiz:** Internet access is required, emailing and chatting is strictly forbidden. Mobile phones should be switched off. R document folder on the ATENEA are allowed during the exam

**Duration:** 2h 00 min

**Marks**: Before 14/Nov/22 Subject ATENEA WEB site.

**Open Office**: Email requests.

**Problem 1: All qüestions account for 1 point (you have to answer 10 out of 15)**

Specifications are given for 506 house related features representing aggregated home data from various suburbs in Boston, Massachusetts **(Boston Housing Data, 1978 -** 1993). The variables recorded include 13 continuous attributes and 2 categorical attributes in **df data frame. Used f data frame**:

|  |  |  |
| --- | --- | --- |
| 1 | **crim** | per capita crime rate by town |
| 2 | **zn** | proportion of residential land zoned for lots over 25,000 sq ft |
| 3 | **indus** | proportion of non-retail business acres per town |
| 4 | **chas** | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| 5 | **nox** | nitric oxides concentration (parts per 10 million) |
| 6 | **rm** | average number of rooms per dwelling |
| 7 | **age** | proportion of owner-occupied units built prior to 1940 |
| 8 | **dis** | weighted distances to five Boston employment centres |
| 9 | **rad** | index of accessibility to radial highways |
| 10 | **tax** | full-value property-tax rate per $10,000 |
| 11 | **ptratio** | pupil-teacher ratio in primary and secondary schools in the neighborhood |
| 12 | **b** | 1000(Bk – 0.63)^2 where Bk is the proportion of afroamericans by town |
| 13 | **lstat** | % lower status of the population - homeowners in the neighborhood considered "lower class" (working poor) |
| 14 | **medv** | Median value of owner-occupied homes in $1000's |
| 15 | **f.hcla** | An additional factor defined by a hierarchical clustering class has been included |

***SOURCE:***

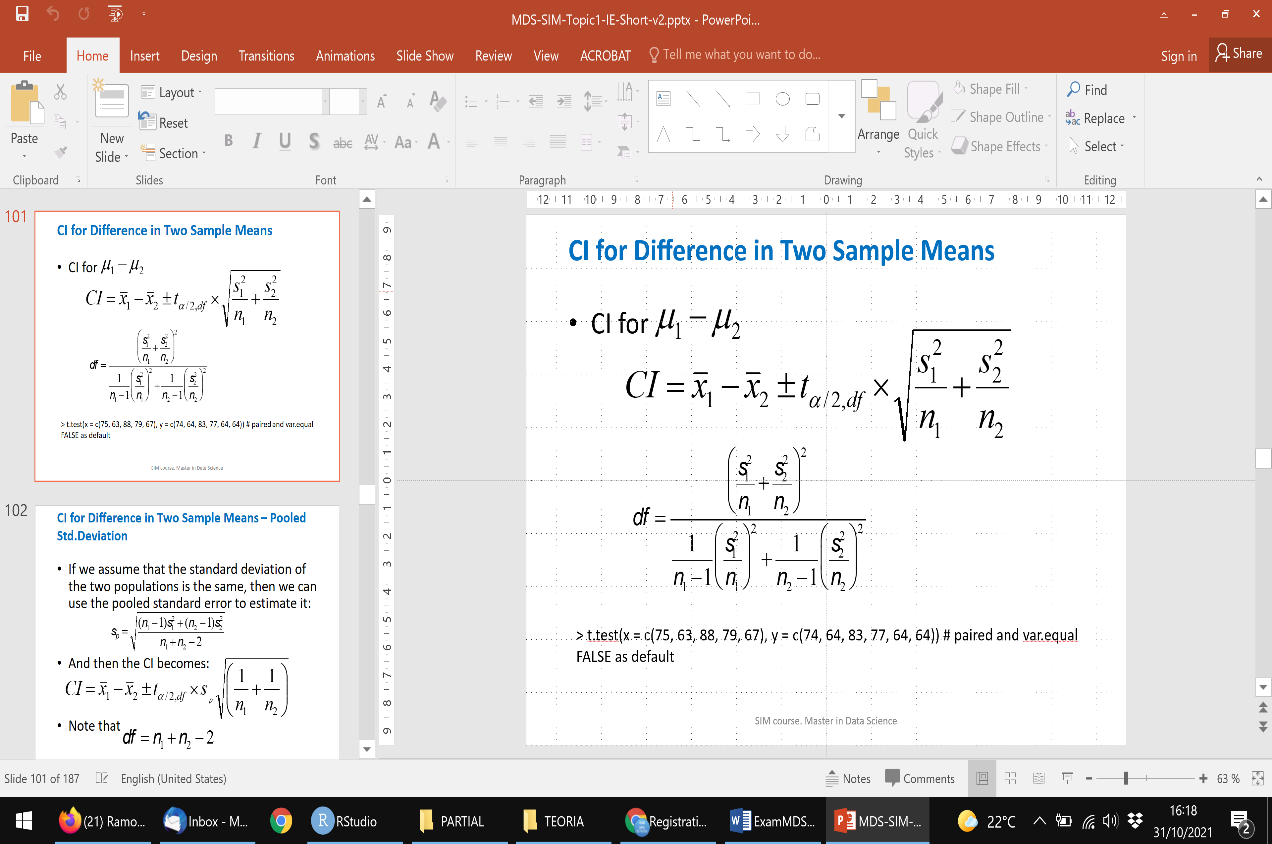
* Origin: This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University and the current version is obtained at **UC Irvine Machine Learning Repository (**https://archive.ics.uci.edu/ml/machine-learning-databases/housing/)**.**
* Creator: Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management,vol.5, 81-102, 1978.

**Median value of owner-occupied homes in $1000's (medv) is going to be our numeric target and chas our target factor when needed.**

1. Some observations have an 'medv' value of 50.0. These data points contain **missing or censored values. Since medv is a numeric target, which suitable actions are needed** before starting a deeper analysis? Implement those actions in your dataset.
2. Determine thresholds for mild and severe outliers for the average number of rooms among homes in the neighborhood. Are there any outliers? Indicate observation id’s and atypical number/s for average rooms.
3. Replace by NA those outliers in RM variable detected in Point 2 and use an imputation procedure discussed in class to fill outlier data points. Assess the consistency of imputed value/s. ***Remove from dataset those observations with NA in RM variable (room number).***

**Calculate the correlation matrix for numerical variables rm, lstat, ptratio and medv.**

1. Would you expect a neighborhood that has an 'LSTAT' value (percent of lower class workers) of 15 have home prices greater or less than those in a neighborhood having a 20 'LSTAT' value?
2. Analyse the profile of the numeric target (medv) using condes() method. A detailed explanation of procedure results is requested.
3. Analyse the profile of the binary target (chas) using a suitable method. A detailed explanation of procedure results is requested.
4. Discuss whether a normal distribution would be a reasonable distribution for medv target.
5. Is there variance homogeneity in the medv target groups defined by f.hcla clusters?
6. Mean medv target can be considered to be the equal across groups defined by f.hcla cluster? Use a two.sided test at 99% confidence.
7. State and test one.sided hypothesis to assess whether medv is greater for f.hclas 1 than for class 3 or the opposite at 99% confidence.
8. The standard deviation of medv in f.hcla 1 should not exceed 10,000$. For the sample in f.hcla 1 in your dataset, calculate the deviation of medv assuming a normal distribution. Stating any assumptions, you need (write them), test at the 1% level the hypothesis that the population standard deviation is larger than 10,000$.
9. Figure out the 99% upper threshold for medv in f.hcla 1 population variance. Normal distribution for medv is assumed to hold.
10. Build a 99% two-sided confidence interval for the difference in the mean of medv between f.hcla 1 and 3. Assume that equal variances in the population medv does not hold and normal distribution of medv (to simplify the calculations), but justify if these assumptions are critical.
11. Determine a 99% confidence interval for the population proportion that favors Riverside in front of Otherwise. Test the null hypothesis that selecting Riverside and Otherwise zones has equal probability.
12. A new survey considered 300 people, 110 prefer Riverside to Otherwise locations. Determine a 99% confidence interval for the difference in the population proportion that favors Riverside in front of other areas accounting the two sources. Test the null hypothesis that selecting Riverside zones has a lower probability in the survey.

Hint:

