Stockholm University

HOME ASSIGNMENT
Statistics and Data Analysis
Department of statistics

You should write a report of **maximum 10 pages** and it should be carried out in work groups of three or four students.

1. The report should be submitted as a PDF file. The file should be named `Asg1grpXwrkYY` where `X` is your group letter and `YY` is your assignments' group number.
2. On the first page, write each group members' name, birthday and email address.
3. Upload the file to Athena: go to Resources/Part 1/Assignment, select the Group that your assignments' group belongs to (A, B, C) and submit your file.
4. Double check that your file was properly submitted.
5. From each group, only one group member is required to submit the assignment.
6. R code should be submitted as a separate file.

Note: If you run into problems when attempting to submit, e-mail your assignments' teacher and attach your report to the e-mail before the deadline expires.

The first deadline is **Sunday, November 24, 23:59**. The second and final acceptance date is **Sunday, December 08, 23:59**. **Note:** After the second deadline, assignments will no longer be accepted.

Notes The criteria for a passing grade are specified in the course description.

- If your report receives a failing grade at the first submission deadline, you may revise it and re-submit. The revised report must be submitted no later than the second deadline. Feedback on the final version can be expected around five working days later.
- If you miss the first deadline, you have a second chance to hand in your report. However, you will not be given a third chance. If you miss the first deadline, and your report receives a failing grade at the second deadline, you will not have the opportunity to revise and re-submit.
- All parts of the assignment must receive a passing grade during the current semester for students to receive credits for the assignment. Partial results cannot be transferred to future semesters.

Note: The dataset needed for exercises 1 to 6 can be found in the folder "Resources/Part 1/Assignment/Datasets" on Athena. The folder contains several Excel files with different sets of data. Use the one that has been assigned to your group.

A real estate agency working in the Stockholm County (Stockholms län) collected data about the housing units they have sold. However, they do not know anything about statistics, therefore they have contacted you and your group asking to write a statistical report that summarizes the collected data, making strong emphasis in the fact that the report should be written in a way that it is understandable for someone who is not familiar with statistics: calculating numbers is something that they could do for themselves, they are mainly interested in a detailed interpretation of all the results.

The dataset contains the following variables:

- *ID:* An identifier of each housing unit;
- *REGION:* The region, within the county, in which the housing unit is located: Northeast, Northwest, Southeast, West or Stockholm (see below for more details);
- *TYPE:* The type of housing unit: Apartment (Lägenhet), Terraced house (Radhus) or Villa;
- *ROOMS:* The number of rooms in the housing unit;
- *AREA:* The size of the housing unit (in square meters);
- *BALCONY:* Indicates whether the housing unit has a balcony or not;
- *STARTING_PRICE:* The price at which the housing unit is initially published (in SEK).

The regions correspond to groups of municipalities as follows:

Northeast: Danderyd, Lidingö, Nacka, Täby, Vallentuna, Upplands Väsby and Järfälla;

Northwest: Solna, Sollentuna and Sundbyberg;

Southeast: Tyresö, Värmdö, Huddinge and Haninge;

West: Södertälje, Nynäshamn, Botkyrka, Salem, Ekerö, Upplands-Bro, Sigtuna, Norrtälje and Österåker;

Stockholm: Stockholm.

Figure 1 shows a map of the different regions.

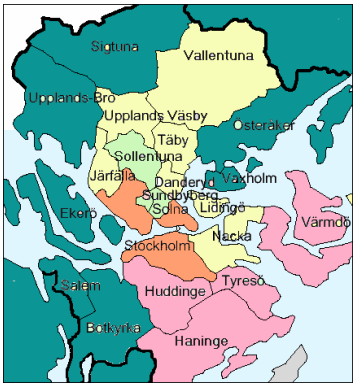

Figure 1: Map of the five regions in which Stockholm County was subdivided: region Northeast in yellow, region Northwest in light green, region Southeast in pink, region West in dark green and Stockholm in orange.

1. Analyze the variable `STARTING_PRICE`. (**Tips: i.** Describe the variable in terms of its location, variability and shape; **ii.** Make use of appropriate numerical and graphical tools; **iii.** Check for the presence of outliers.)

2. Analyze simultaneously the variables `REGION` and `TYPE`. (**Tips: i.** Make use of appropriate numerical and graphical tools; **ii.** The real estate agency may be particularly interested in knowing if the type of housing units differ between regions.)

3. Analyze simultaneously the variables `REGION` and `AREA`. (**Tips: i.** Make use of appropriate numerical and graphical tools; **ii.** The real estate agency may be particularly interested in knowing if the size of the housing units differ between regions.)

4. Analyze the relationship between the variables `STARTING_PRICE` and `AREA`. (**Tips: i.** Make use of appropriate numerical and graphical tools; **ii.** Take into account that the price can be considered to be dependent on the area.)

5. Fit a linear regression that explains `STARTING_PRICE` in terms of `AREA`. (**Tips: i.** Make use of appropriate numerical and graphical tools; **ii.** Make sure that you report the estimated regression in equation form and graphically; **iii.** Make sure to indicate how good is the area at explaining the price; **iv.** Make sure you interpret the coefficients of the regression.)

6. Fit a linear regression that explains `STARTING_PRICE` in terms of `REGION`, `TYPE`, `BALCONY`, `ROOMS` and `AREA`. (**Tips: i.** Make use of appropriate numerical and graphical tools; **ii.** Make sure that you report the estimated regression in equation form; **iii.** Make sure to indicate how good are the independent variables at explaining the price; **iv.** Make sure you interpret the coefficients of the regression.)

7. The dataset `test.xlsx` that can be found in Athena (Resources/Part 1/Assignment) contains information on the region, the type of housing unit, the presence of a balcony, the number of rooms and the area of ten housing units. Use your regression model from exercise 6 to predict the starting price of these housing units.