Assignment 1

Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

Data Analysis 2

2024/2025 Fall

Your task is to analyze the pattern of association between Airbnb prices and the features of the place (which you can select). The aim of this assignment is to guide you through a simple regression analysis on live data. This assignment is evaluated for good analysis (statistical argument) only.

You need to upload a pdf file containing the analysis on ceulearning site.

Selecting the city deadline: Sunday, 10th of November 2024, 23:59
Google spreadsheet – failing to submit will result in -1p / day
Final deadline: Sunday, 24 November 2024, 23:59, on ceulearning

1 Task

1. Check the Airbnb data, which is collected by Inside Airbnb a community supporting data and advocacy about Airbnb's impact on residential communities.
   (a) Always obtain information on the data you are using! This data was collected and released by Airbnb as part of or insideAirbnb `http://insideairbnb.com/get-the-data.html`. The data behind the Inside Airbnb website are sourced from publicly available information from the Airbnb website. The data have been analyzed, cleaned, and aggregated where appropriate to facilitate public discussion. Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license.
2. Sign up for google spreadsheet where each of you chooses their city. **Selecting the city deadline: Sunday 10th November 2024, 23:59** You should select a city that nobody used from your cohort (see spreadsheet).
   (a) You have to select your favorite city from the available 116 cities. See the Get Data section from Inside Airbnb.
   (b) Make sure you have the price variable and a continuous variable (such as distance or rating).
3. **Cleaning the data**
   (a) Be curious and check all your variables!
   (b) Select variables that you find useful (at least some that are available and listed below) and convert them into the appropriate format (numeric or factor)
      i. You must have `id, price` and your favorite continuous variable.
      ii. Do not forget entity resolution, handling duplicates and missing observations.
   (c) If your data has any of the following variables, keep them and filter them consciously. Write a comment about what and why you filter and also mention it **briefly** in the report as well
      i. `property_type`, `room_type`, `cancellation_policy`, `bed_type`, `neighbourhood_cleansed`,

        ii. or any other variable that you might find important

   (d) Check your main variables: price and your selected continuous variable

      i. Filter if needed (e.g. duplicates, errors, etc.)

      ii. Create descriptive statistics

      iii. Check the histograms

      iv. it is good practice to make yourself comments: why have you decided for certain filtering and if there are possible nonlinearities that you want to model later on

   (e) Make sure you have **at least 200 observations** after cleaning the data

4. **Analysis of the data**

   (a) Use price as the dependent variable (y) and your continuous variable as x. Note that this is only a 'potential' variable, thus you may transform it with ln transformation, piecewise linear spline, polynomial or any other type of reasonable nonlinear transformation (see later point)

   (b) Check both your variables – with the help of histograms, summary statistics and checking extreme values – and make a conscious decision on which observation(s) to drop.

   (c) Choose a proper scaling for your variable (may divide by a thousand, a million, ect.). Stay consistent during the interpretations with the scaling!

   (d) Check your distributions for y and x variables: use histograms and summary statistics table (mean, median, min, max, P05, P75, standard deviation, and number of obervations)

   (e) Check the possible different ln transformation for the variables by plotting different scatter plots with lo(w)ess. **Make substantive and statistical reasoning**, where and when to use ln transformation. You do not need to fit any model here, only use substantive/statistical reasoning based on the graphs.

      i. Take care when making the ln transformation: you may need to drop or shift some observations and include another dummy variable.

   (f) Choose your specification for the nonlinear transformation and estimate the following models with graphical visualizations:

      i. Simple linear regression with continuous variable only

      ii. Most reasonable model using nonlinear transformation(s)

      iii. Note: always use heteroscedastic SEs.

   (g) Compare the models and choose your preferred one

      i. Use substantive and statistical reasoning for your chosen model.

      ii. Show the model results in the report along with the most telling graph.

      iii. Report the model comparison (and all other estimated model results) **in the appendix**.

   (h) You need to test your β parameter for your continuous explanatory variable. (In case of quadratic test β_1 that is the linear component/first slope.)

      i. Carry out the following test: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$.

   (i) Finally, using your selected model, analyze the residuals:

      i. Find 5 places with the lowest prices compared to its predicted value

      ii. Find 5 places with the highest prices compared to its predicted value

5. **Report**

   (a) Create sections:

      i. Executive summary

      ii. Introduction

      iii. Data

      iv. Model and evaluation

      v. Conclusion

   (b) We will evaluate only **pdf** report. Make sure you have done all the main points and select the needed parts in the report. Put everything else into the appendix.

2 Evaluation

Overall you can earn 10 points.

1. Executive summary (1p)
   (a) 2 sentences about the main results of the analysis: what is the variable, what does the pattern of association look like, what model do you use, what is the main message of your model.
2. Introduction: aim of the analysis and introduction of your variables (1p)
   (a) What you want to show in this analysis, 1 sentence - (0.5/1p)
   (b) What are your variables and how is it measured? What is the population and how does your sample relate to that? What are the potential data quality issues? 2-3 sentences - (0.5/1p)
3. Data - (3p)
   (a) Selecting observations (or dropping) and potential scaling of them. 2 sentence - (1/3p)
   (b) Summary statistics for X and Y with 2-3 sentences, explain the main features and histogram if needed. - (1/3p)
   (c) Investigate the transformation of your variables (You can put your graphs in the appendix, but not into the main part.) - (1/3p)
      i. Substantive reasoning 2-3 sentences
      ii. Statistical reasoning with the help of the descriptive stats/graphs, 1-2 sentences
4. Model and Evaluation (4p)
   (a) Estimating different models (1/4p)
      i. Model comparison table with scatter plot visualization in the appendix with explanations of what you can see in the table. 5-8 sentences
      ii. State your choice of model: substantive and statistical reasoning. 3-5 sentences in the appendix
   (b) Presentation of your model choice (1/4p)
      i. State your choice of model in an appropriate format. 1 formula
      ii. Interpret the estimated parameters of the model. 2-3 sentence
   (c) Hypothesis testing on β (which interacts with X) - (1/4p)
      i. Test the following hypothesis: $H_0 : \beta = 0, \quad H_A : \beta \neq 0$.
      ii. Argue in one sentence for your choice of SE.
      iii. Choose a significance level and draw your conclusions.
      iv. Report your results in a table and write 1-3 sentences in the conclusion.
   (d) Analysis of the residuals (1/4p)
      i. Interpretation for places with the largest negative errors. 2-3 sentences
      ii. Interpretation for places with the largest positive errors. 2-3 sentences
5. Formatting of your report (1p)
   (a) How does the document look: graphs and tables are readable, report is well formatted.
   (b) Your main text (non-appendix sections) in a pdf should not exceed 4 pages:
      i. Each extra page is penalized by 0.5 points.
   (c) Your appendix can be as long as you wish. You should put the appendix in the same document, and indicate it with a new section or title.

Good luck!