Lab 1: Exploratory Data Analysis of CEO Salary Data

w203 Instructional Team

May 20, 2017

Introduction

In this exercise, imagine that you (and your team) are hired by a think tank that is preparing a report on CEO salary. The think tank is especially interested in whether company performance is related to salary. They have collected data on a selection of companies, provided in the file ceosal_w203.RData. Just like numerous data science problems, which are originated from and/or motivated by real world business, policy, or scientific questions, the think tank only give you very limited information. In fact, that's all they tell you. Luckily, they also give you a very brief data dictionary (or a codebook):

1. salary	1990 compensation, $1000s
2. age	in years
3. college	=1 if attended college
4. grad	=1 if attended graduate school
5. comten	years with company
6. ceoten	years as ceo with company
7. profits	1990 profits, millions
8. mktval	market value, end 1990, mills.

Note that this is a real data set, but it has been modified by the instructors.

You are to conduct an exploratory data analysis of this dataset to address the think tank's questions. Remember that to use descriptive statistical analysis tools and note any features you find that you think would be relevant to (subsequent) statistical modeling.

Deliverables

You are to work in a group of 2 or 3 students. Only one student in the group needs to submit via the ISVC, but make sure that you include the names of all group members in your report.

You must turn in

1. Your pdf report. In this report, do not suppress the R code that generates your output.
2. The R script you use to generate your report

For example, most students will find it convenient to write an Rmd file and turn it in along with the pdf it generates.

Be sure to follow the guidelines we covered in class. In particular, do not include any output that you do not discuss in your text.

Please include a brief introduction and a brief discussion that explains your findings at a high level.

Your analysis should be thorough, but limit your report to a maximum of 25 pages. This means that you will have to make choices about what variables and relationships to focus on (and justify those choices).

Due Date

This lab is due 24 hours before the week 4 live session. It would be **June 5th** for the Tuesday session.

General Assessment

Instructors will use the following rubric as a general guide for grading your work. Please note that a report with simply an "output-dump", a practice that justs "dump" a ton to graphs and tables without much narrative and explanation, will receive a very low score.

- Effective intro and conclusion - 20 points
- Thorough univariate analysis of each key variable - 20 points
- Clear understanding of key bivariate relationships - 30 points
- Exploration of secondary variables - 10 points
- Analysis of potential confounding effects - 10 points
- Handling of variable coding issues and missing values - 10 points
- Other deductions - other errors will result in deductions according to instructor judgement