

# Exploratory Data Analysis

Part 1, session 1a of Data Mining Intro

2025-01-19 07:30 GMT

## Abstract

Review concepts and techniques of exploratory data analysis.

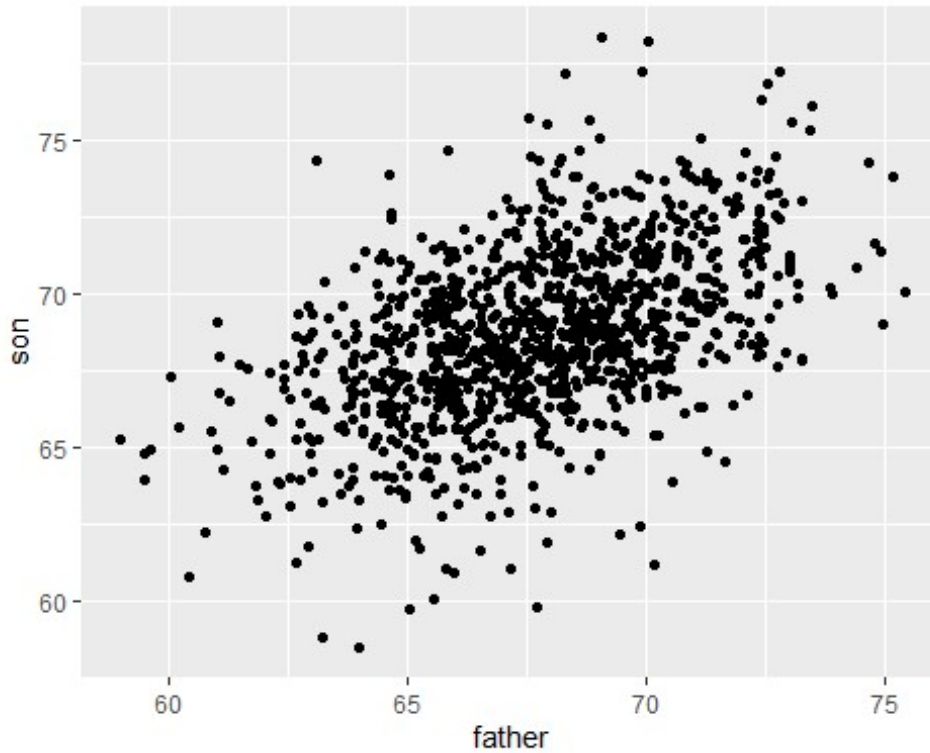
## Table of Contents

Heights of Fathers and Sons .....	1
Class Exercise: Diamond Data .....	3
Class Exercise: General Social Survey Data .....	4
Discussion: what should EDA mean? .....	4
Team Exercises .....	6
Resources.....	7

---

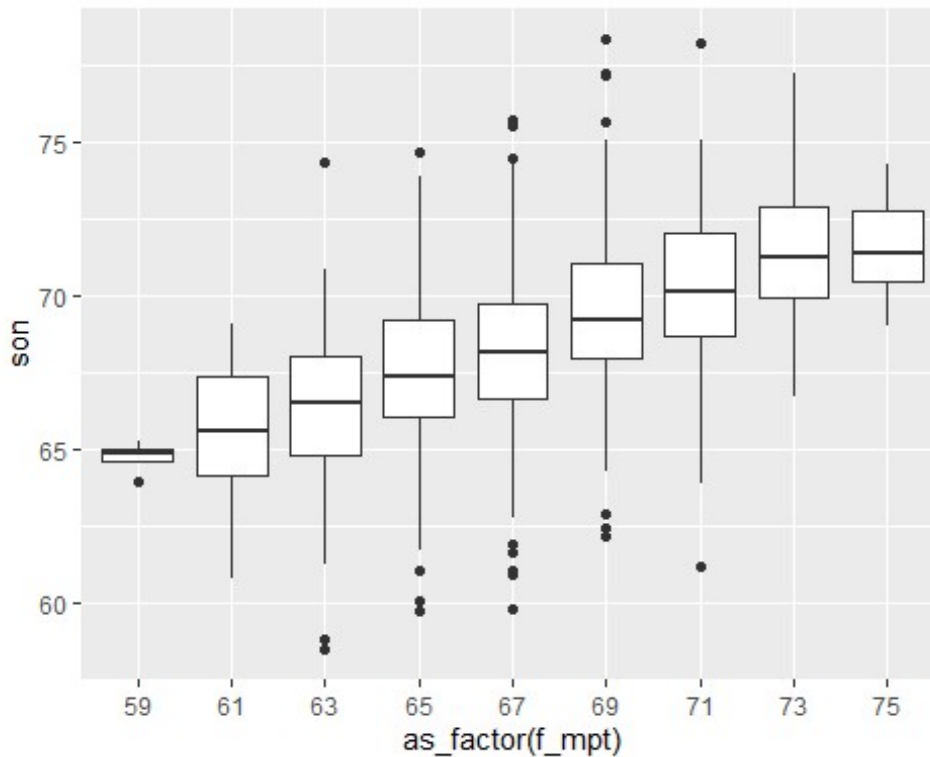
## Heights of Fathers and Sons

In 1885 Sir Francis Galton examined the heights of parents and their children to determine the strength of evidence to support height as a hereditary trait. Galton's subsequent protege, Karl Pearson (who went on in 1911 to form the world's first statistics department at University College London) followed up with an analysis of the heights of fathers and sons.



In the scatter diagram above, each point represents the respective heights of a father-son pair. The density of points in the middle range of heights obscures individual points.

Here is another view of the same data obtained by grouping father heights into successive intervals (each one labeled by its mid-point) and then constructing a box-plot of sons' heights within each interval.



```
## # A tibble: 9 × 6
## # Groups:   f_ivl [9]
##   f_ivl  f_mpt s_count s_min s_max s_avg
##   <fct> <dbl>   <int> <dbl> <dbl> <dbl>
## 1 (58,60]    59     4  63.9  65.2  64.7
## 2 (60,62]    61    16  60.8  69.1  65.5
## 3 (62,64]    63    77  58.5  74.3  66.3
## 4 (64,66]    65   208  59.8  74.7  67.5
## 5 (66,68]    67   276  59.8  75.7  68.2
## 6 (68,70]    69   275  62.2  78.4  69.5
## 7 (70,72]    71   152  61.2  78.2  70.2
## 8 (72,74]    73    63  66.7  77.2  71.4
## 9 (74,76]    75     7  69.0  74.3  71.6
```

Galton noted that in general sons were taller than fathers, with extremely tall or short fathers corresponding to not quite so extremely tall or short sons, respectively. This observation led Galton to coin the term “regression toward mediocrity”, or (more politely) “regression toward the mean”.

## Class Exercise: Diamond Data

Team up with a classmate and load the diamond data provided by R package ggplot2. How many rows of data are there? How many columns? What questions occur to you about the data? How might you address those questions? Take 15 minutes to prepare to report out to the class.

## Class Exercise: General Social Survey Data

The [General Social Survey](#) is a long-running US survey conducted by the independent research organization NORC at the University of Chicago. The survey consists of thousands of questions designed to monitor changes in social characteristics and attitudes in the US. In R package `forcats`, Hadley Wickham formed the data set `gss_cat` consisting of just a few columns to illustrate the challenges one encounters when working with categorical (non-numeric) variables (“factors”).

Team up with a classmate and load the `gss_cat` data-set provided by R package `forcats`. How many rows of data are there? How many columns? What questions occur to you about the data? How might you address those questions? Take 15 minutes to prepare to report out to the class.

## Discussion: what should EDA mean?

Exploratory Data Analysis (EDA) is an approach to data analysis advocated by [John Tukey](#), a leading American statistician of the 20th century. The approach contrasts with what Tukey called “confirmatory analysis”, that is, a focus on probability models of data-generation along with the estimation or testing of model parameters. The difference is one of emphasis: EDA includes models suggested by data, but with an emphasis on understanding current and potential data sets.

The exploration is led by one’s questions about the data. Relevant questions may or may not be obvious (or given). Variables may or may not be readily categorized as “response variables” versus “predictor variables”. The ability to develop and recognize relevant questions is an important skill largely gained through experience.

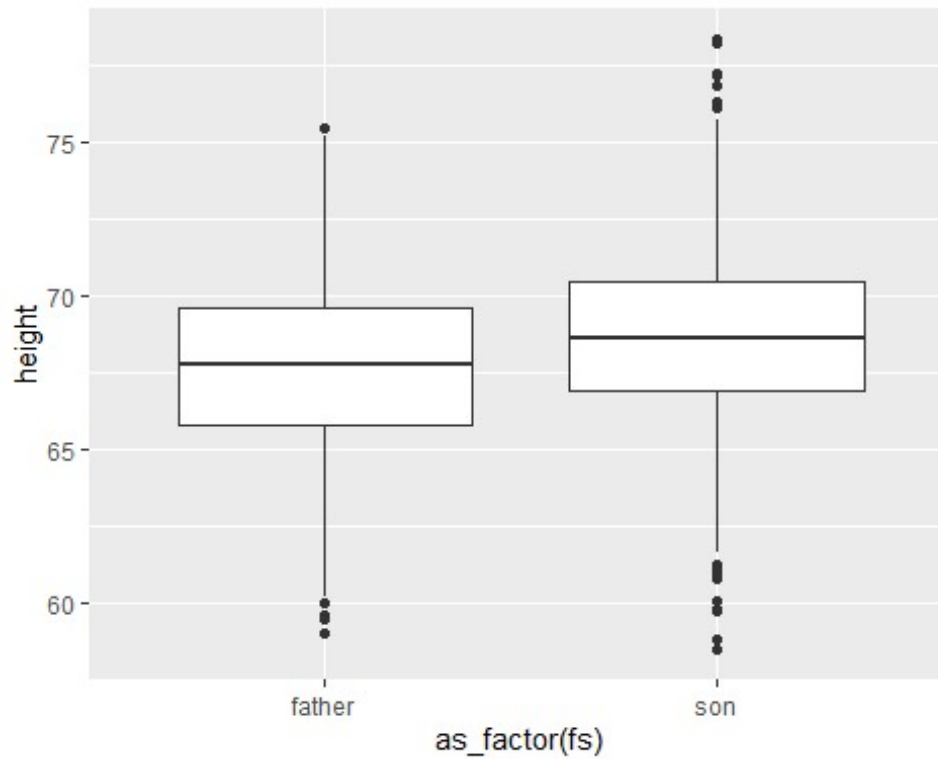
Important EDA outcomes include

- the discovery of unanticipated data patterns, and
- proposals to examine tentative answers suggested by the current data, perhaps using a new data set designed for this purpose.

EDA methods are used within the context of confirmatory analysis to examine the data for errors not encompassed by the models under study (e.g., errors in data transcription or transmission), or to search for other departures from model assumptions.

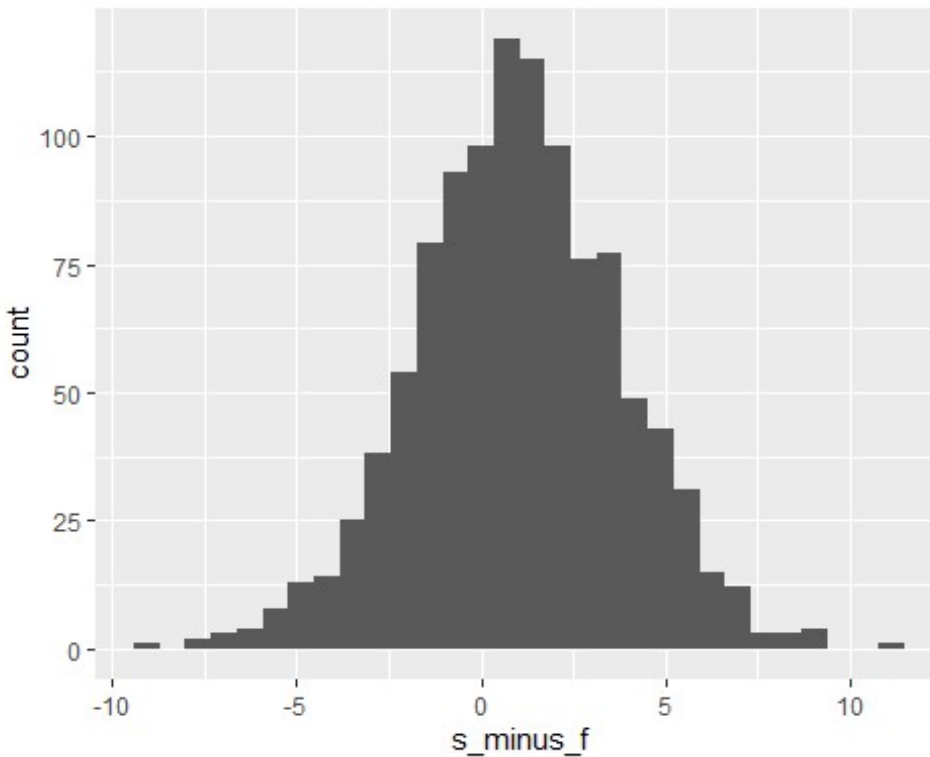
EDA methods can be broadly understood as the methods of descriptive statistics: data summaries (graphical or tabular) intended to enhance our understanding of the data. EDA differs from descriptive statistics in a reliance on the questions of the data analyst and a readiness to examine various transformations of the data.

As an example, here are some ways we might address the question of how, if at all, the heights of fathers and sons differ in the data presented above.



```
## # A tibble: 2 × 5
##   fs      min  mid  avg  max
##   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 father  59.0  67.8  67.7  75.4
## 2 son    58.5  68.6  68.7  78.4
```

The box-plot and table above show that, on average, sons are about an inch taller than fathers. Here's a histogram of the difference in heights (son minus father) across father-son pairs.



##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-9.0	-0.9	1.0	1.0	2.7	11.2

The summary of individual differences in height (son minus father) strengthens the previous aggregate summaries: the distribution of son's height minus father's height is fairly symmetric around a difference of about one inch.

## Team Exercises

1. Response versus predictor variables: for each of the data sets presented above, propose one or more variables as response variables. How, if at all, might someone else argue for a different choice? What should we mean by “response” and “predictor” variables? Describe a situation in which this distinction would not be suitable.
2. Regression to the mean: the discussion of father-son heights uses the phrase “extremely tall or short fathers corresponding to not quite so extremely tall or short sons, respectively”. How would you formulate the meaning of that phrase? Do the data demonstrate this phenomenon?
3. Diamond data: propose a question for the diamond data, and then try to address that question. What (if anything) did you learn from this task?
4. Survey data: propose a question for the gss\_cat data, and then try to address that question. What (if anything) did you learn from this task?

## Resources

[simpleR](#) Using R for Introductory Statistics.

By [John Verzani](#) Department of Mathematics at CSI

[R for Data Science \(2e\)](#), by [Hadley Wickham](#)

RPubs - [Diamonds Dataset](#)

NORC's General Social Survey ([GSS](#))

[EDA](#): Exploratory data analysis - Wikipedia