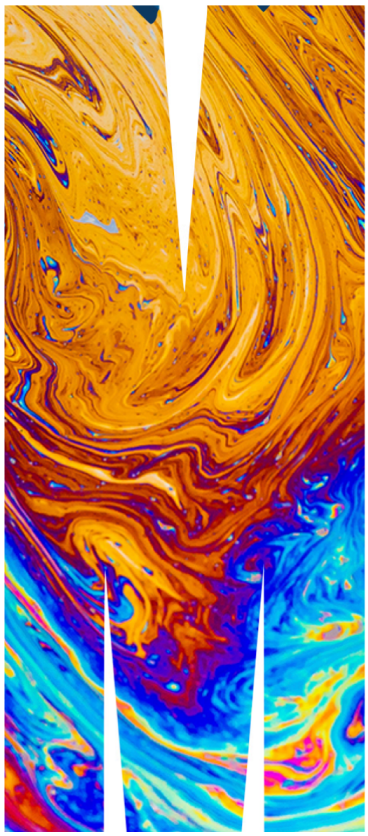
MONASH University

## ETC5521: Exploratory Data Analysis

### Introduction

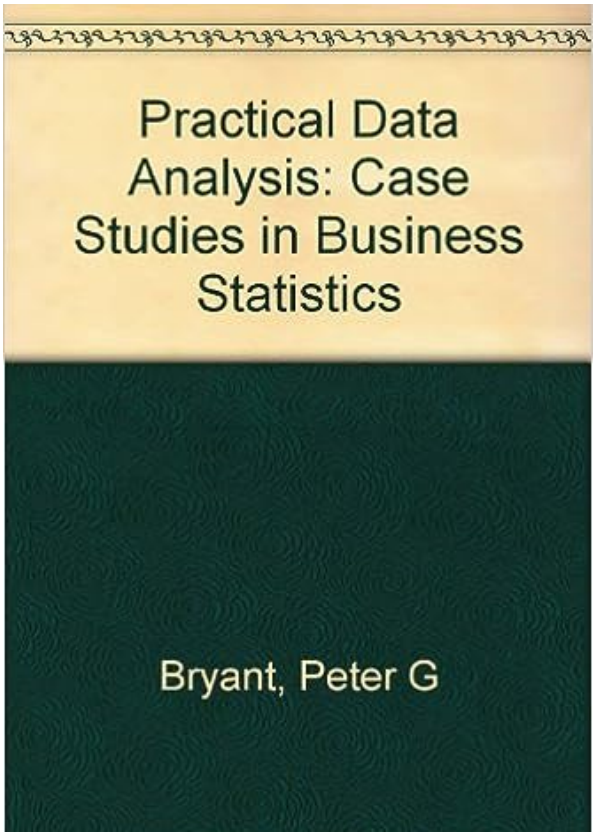


Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 1 - Session 2

# A simple example to illustrate "exploratory data analysis" contrasted with a "confirmatory data analysis"




### What are the factors that affect tipping behaviour?

In one restaurant, a food server recorded the following data on all customers they served during an interval of two and a half months in early 1990.

Food servers' tips in restaurants may be influenced by many factors, including the nature of the restaurant, size of the party, and table locations in the restaurant. Restaurant managers need to know which factors matter when they assign tables to food servers.

| Variable | Explanation |
|---|---|
| obs | Observation number |
| totbill | Total bill (cost of the meal), including tax, in US dollars |
| tip | Tip (gratuity) in US dollars |
| sex | Sex of person paying for the meal (0=male, 1=female) |
| smoker | Smoker in party? (0=No, 1=Yes) |
| day | 3=Thur, 4=Fri, 5=Sat, 6=Sun |
| time | 0=Day, 1=Night |
| size | Size of the party |



# General strategy for EXPLORATORY DATA ANALYSIS

Its a good idea to examine the data description, and the explanation of the variables.

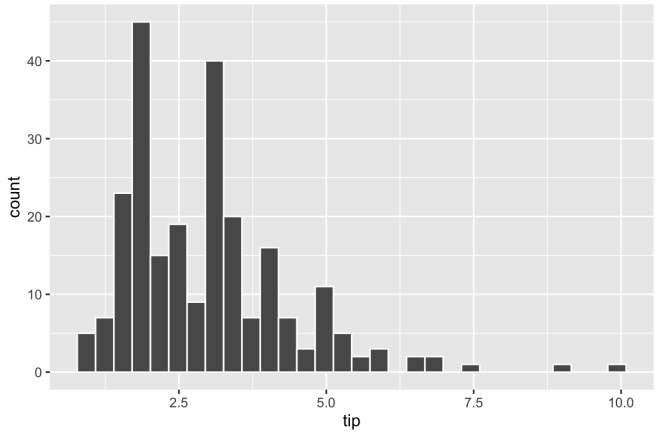
> You need to know what type of variables are in the data in order to decide appropriate choice of plots, and calculations to make.
>
> Data description should have information about data collection methods, so that the extent of what we learn from the data might apply to new data.

What does that look like here?

```
## # A tibble: 1 × 8
##     obs totbill   tip sex   smoker day   time    size
##   <dbl>   <dbl> <dbl> <chr> <chr>  <chr> <chr> <dbl>
## 1     1    17.0  1.01 F     No     Sun   Night     2
```

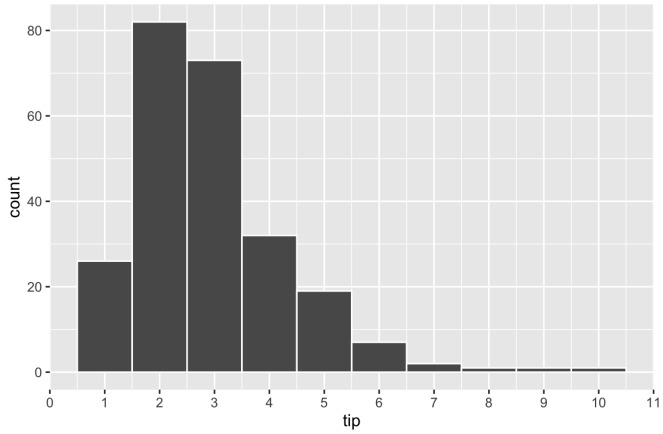
Look at the distribution of quantitative variables tips, total bill.
Examine the distributions across categorical variables.
Examine quantitative variables relative to categorical variables

```
ggplot(tips,
    aes(x=tip)) +
  geom_histogram(
    colour="white")
```





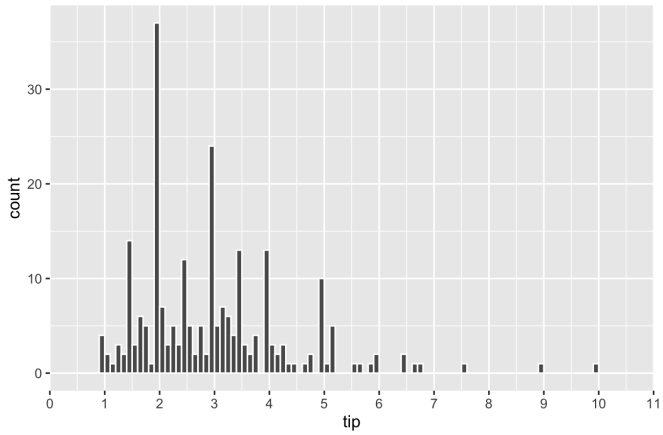

```
ggplot(tips,
    aes(x=tip)) +
  geom_histogram(
    breaks=seq(0.5,10.5,1),
    colour="white") +
  scale_x_continuous(
    breaks=seq(0,11,1))
```




Big fat bins. Tips are skewed, which means most tips are relatively small.



```
ggplot(tips,
    aes(x=tip)) +
  geom_histogram(
    breaks=seq(0.5,10.5,0.1),
    colour="white") +
  scale_x_continuous(
    breaks=seq(0,11,1))
```




Skinny bins. Tips are multimodal, and occurring at the full dollar and 50c amounts.

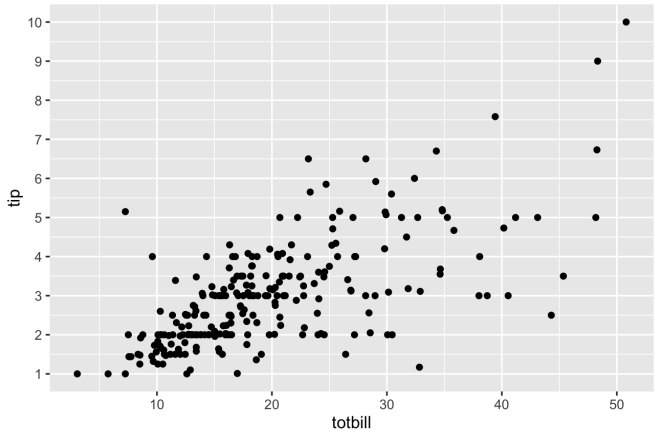
## We could also look at total bill this way

but I've already done this, and we don't learn anything more about the multiple peaks than waht is learned by plotting tips.



## Relationship between tip and total

```
p <- ggplot(tips,
    aes(x= totbill, y=tip)) +
  geom_point() +
  scale_y_continuous(
    breaks=seq(0,11,1))
p
```

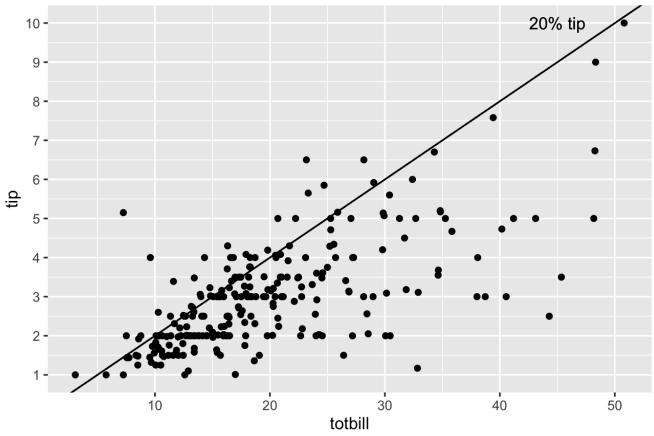



Why is total on the x axis?

Should we add a guideline?

## Add a regression line

```
p <- p + geom_abline(intercept=0,
                     slope=0.2) +
  annotate("text", x=45, y=10,
           label="20% tip")
p
```




Most tips less than 20%: Skin flints vs generous diners

A couple of big tips

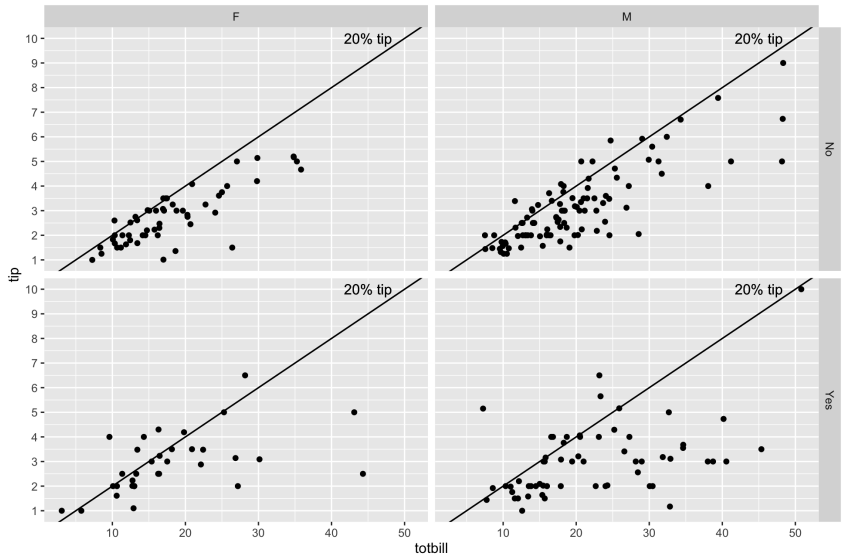
Banding horizontally is the rounding seen previously



## We should examine bar charts and mosaic plots of the categorical variables next

but I've already done that, and there's not too much of interest there.

```
p + facet_grid(smoker~sex)
```






# What do we learn?

The bigger bills tend to be paid by men (and females that smoke).

Except for three diners, female non-smokers are very consistent tippers, probably around 15-18% though.

The variability in the smokers is much higher than for the non-smokers.

**Isn't this interesting?**

## Procedure of EDA

We gained a wealth of insight in a short time.

Using nothing but graphical methods we investigated univariate, bivariate, and multivariate relationships.

We found both global features and local detail. We saw that

- tips were rounded; then we saw the obvious
- correlation between the tip and the size of the bill, noting the scarcity of generous tippers; finally we
- discovered differences in the tipping behavior of male and female smokers and non-smokers.

These are unexpected delights! We would have missed these insights if we had focused solely on the primary question.

## Getting real

The preceding explanations may have given a somewhat misleading impression of the process of data analysis.

The data had no problems; for example, there were no missing values and no recording errors.

Every step was logical and necessary.

Every question we asked had a meaningful answer.

Every plot that was produced was useful and informative.

In actual data analysis, nothing could be further from the truth.

Real datasets are rarely perfect;

Most choices are guided by intuition, knowledge, and judgment;

Most steps lead to dead ends;

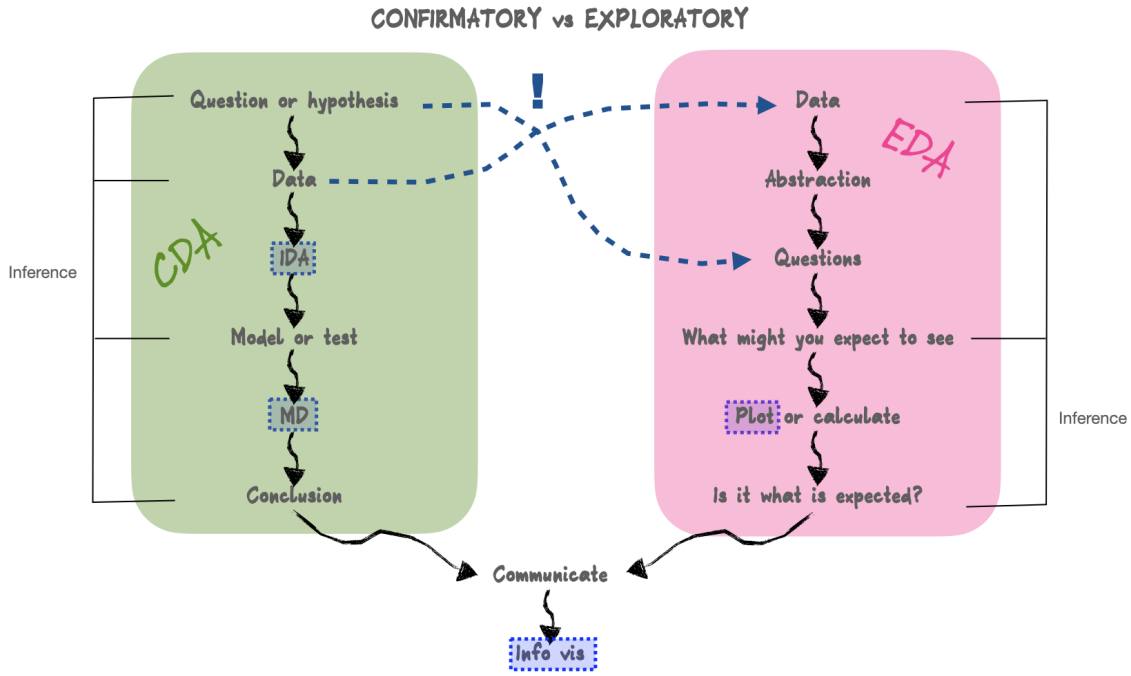
Most plots end up in the wastebasket.

This may sound daunting, but even though data analysis is a highly improvisational activity, it can be given some structure nonetheless.



## Tips example is an illustration only

Although we have focused on the analysis of the tips it only serves the purpose of an example to illustrate the difference between confirmatory and exploratory analysis.






# Exploratory data analysis

Prof Di says:

I like to think of EDA as making time to "play in the sand" to allow us to find the unexpected, and to some better understand our data.

We like to think of this as a little like traveling. We may have a purpose in visiting a new city, perhaps to attend a conference, but we need to take care of our basic necessities, such as finding eating places and gas stations. Some of our movements will be pre-determined, or guided by the advice of others, but some of the time we wander around by ourselves. We may find a cafe we particularly like or a cheaper gas station. This is all about getting to know the neighborhood.

EDA has always depended heavily on graphics, even before the term data visualization was coined. A favorite quote from John Tukey's rich legacy is that we need good pictures to *"force the unexpected upon us."*

# What can go wrong?

## Is it data snooping?

Because EDA is very graphical, it sometimes gives rise to a suspicion that patterns in the data are being detected and reported that are not really there. (Stay tuned, we'll provide solutions later in the semester.)

So many different combinations may be examined, that something is bound to be interesting.

! It is a problem if structure seen in the plot drives hypothesis testing on same data.

Sometimes this is called data snooping.

An abuse of exploration happens when data is modified, typically after examining it, to achieve a significant $p$-value. For example, observations might be dropped, or some data processing made. This is called p-hacking, and it is unethical.

Or, many comparisons are made, but only the significant ones are reported.

# In defense of EDA

We snooped into the tips data, and from a few plots we learned an enormous amount of information about tipping:

There is a scarcity of generous tippers,

the variability in tips increases extraordinarily for smoking parties, and

people tend to round their tips.

These are very different types of tipping behaviors than what we learned from the regression model. The regression model was not compromised by what we learned from graphics, and indeed,

we have a richer and more informative analysis. Making plots of the data is just smart.



Words of wisdom

*False discovery is the lesser danger when compared to non-discovery. Non-discovery is the failure to identify meaningful structure, and it may result in false or incomplete modeling. In a healthy scientific enterprise, the fear of non-discovery should be at least as great as the fear of false discovery.*

## Why aren't there more courses on EDA?

> Teaching data analysis is not easy, and the time allowed is always far from sufficient. But these difficulties have been enhanced by the view that "avoidance of cookbookery and growth of understanding come only by mathematical treatment, with emphasis upon proofs." The problem of cookbookery is not peculiar to data analysis. But the solution of concentrating upon mathematics and proof is.

Tukey 1962 The Future of Data Analysis



## There really are many courses

Every introductory statistics course begins with exploratory data analysis, and teaches box plots. It is just a simple treatment, though.

A book by Peng, and a Coursera class by Peng, Leek and Caffo with more than a 100,000 currently enrolled.

## At Monash

ETC1010/5510 - Introduction to data analysis

ETF5922 - Data visualisation and analytics

FIT3152 - Data analytics

FIT5197 - Modelling for data analysis

FIT5149 - Applied data analysis

FIT5145 - Introduction to data science

FIT5147 - Data exploration and visualisation

STA2216 - Data analysis for science

all have parts that would be considered exploratory data analysis.

You've just completed ETC5510 Introduction to data analysis. Isn't this EDA?

Yes! Think about this course (ETC5521) as advanced exploratory data analysis. We will go a bit deeper, with more structure, and historical background, and venture in with EDA attitude.



## Ready?

# Resources


Cook and Swayne (2007) Interactive and Dynamic Graphics for Data Analysis, Introduction

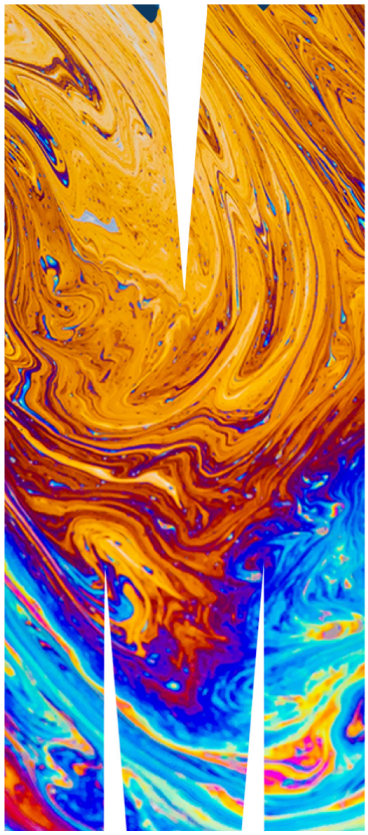
Donoho (2017) 50 Years of Data Science

Staniak and Biecek (2019) The Landscape of R Packages for Automated Exploratory Data Analysis




MONASH University






Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu


Week 1 - Session 2