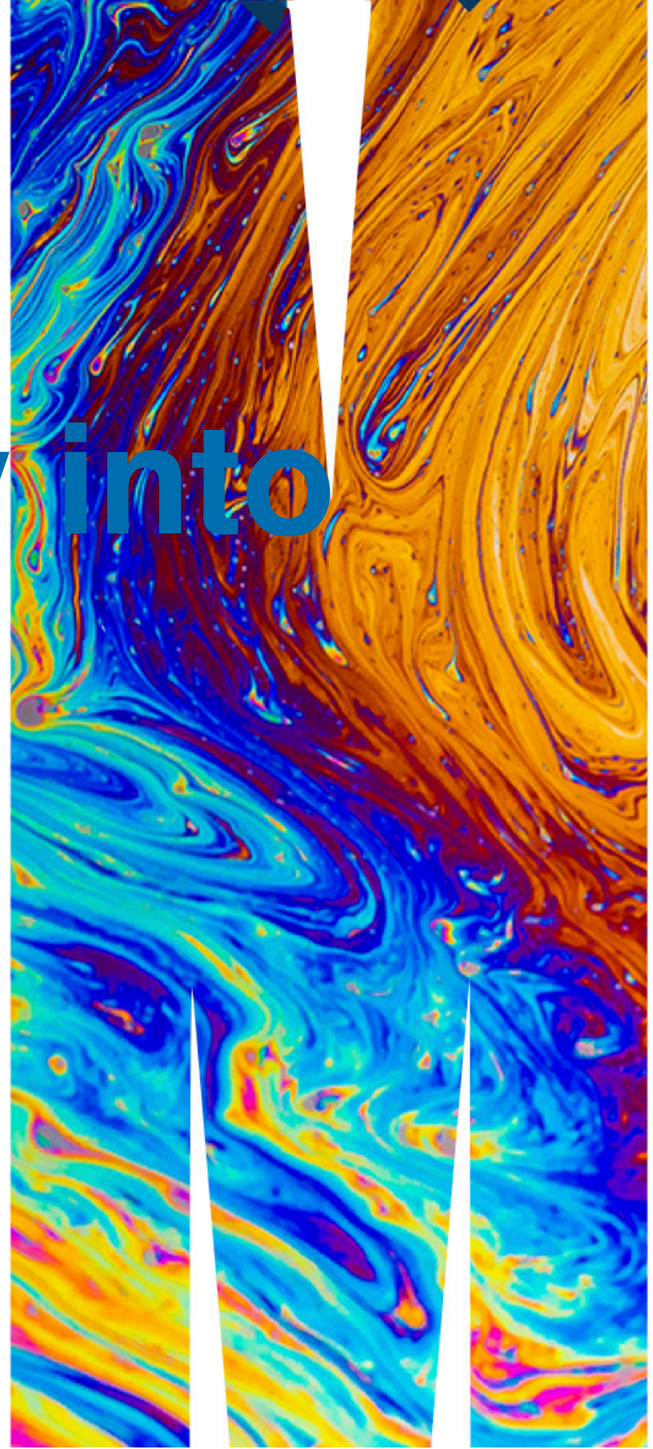
MONASH University

ETC5521: Diving Deeply into Data Exploration

Week 1: Introduction

Professor Di Cook

Department of Econometrics and Business Statistics



About this unit

Teaching team 1/2

Di Cook
Distinguished Professor
Monash University

🌐 https://dicook.org/
✉️ ETC5521.Clayton-x@monash.edu
🦣 @visnut@aus.social

- I have a PhD from Rutgers University, NJ, and a Bachelor of Science from University of New England
- I am a Fellow of the American Statistical Association, elected member of the the R Foundation and International Statistical Institute, Past-Editor of the Journal of Computational and Graphical Statistics, and the R Journal.
- My research is in data visualisation, statistical graphics and computing, with application to sports, ecology and bioinformatics. I likes to develop new methodology and software.
- My students always work on methods and software that is generally useful for the world. They have been responsible for bringing you the tidyverse suite, knitr, plotly, and many other R packages we regularly use.

Teaching team 2/2

Krisanat Anukarnsakulchularp
Master of Business Analytics
Monash University

🌐 https://github.com/KrisanatA

✉️ ETC5521.Clayton-x@monash.edu

- He has a Bachelor of Actuarial Science, Monash University, 2018 - 2021
- and a Master of Business Analytics, Monash University I 2022 - 2023.
- He has published the R package animbook
- and hopes to be a PhD student at Monash from 2025.
- This is his second semester tutoring at Monash, and one of several units working on this semester.

Got a question, or a comment?

🖐 🔡 You can ask directly by unmuting yourself, or typing in the chat, of the live lecture.

💻 If watching the recording, please post in the discussion (ED) forum.

Welcome!

Before modelling and predicting, data should first be explored to uncover the patterns and structures that exist. Exploratory data analysis involves both numerical and visual techniques designed to reveal interesting information that may be hidden in the data. However, an analyst must be cautious not to over-interpret apparent patterns, and to properly assess the results of a data exploration.

1. learn to use modern data exploration tools with many different types of contemporary data to uncover interesting structures, unusual relationships and anomalies.
2. understand how to map out appropriate analyses for a given data set and description, define what we would expect to see in the data, and whether what we see is contrary to expectations.
3. be able to compute null samples in order to test apparent patterns, and to interpret the results using computational methods for statistical inference.
4. critically assess the strength and adequacy of data analysis.

📅 Unit Structure

- 2 hour lecture 👩‍💻 Tue 10.00am - noon, on zoom (see moodle for the link) *Class is more fun if you can attend live!*
- 2 x 1.5 hour on-campus tutorial 🛠️ Wed 9:30-11:00 and Wed 7:30-9:00pm CL_Anc-19.LTB_134 *Attendance is expected - this is the chance to practice what is explained in lecture under your tutor's guidance.*

Resources

Course homepage: this is where you find the course materials (lecture slides, tutorials and tutorial solutions) https://ddde.numbat.space/

Moodle: this is where you find discussion forum, zoom links, and marks https://learning.monash.edu/course/view.php?id=18864

GitHub classroom: this is where you will find assignments, but links to each will be available in moodle. https://classroom.github.com/classrooms/175896553-etc5521-2024-classroom-29a96a

💯 Assessment Part 1/2

- Weekly quizzes (**5%**) There will be a weekly quiz starting week 2 provided through Moodle. These are a great chance to check your knowledge, and help you prepare for the tutorial and to keep up to date with the weekly course material. Your best 10 scores will be used for your final quiz total.
- Assignment 1 (**15%**), through GitHub classroom, Due: Aug 5, 11:55pm. This is an individual assessment.
- Assignment 2 (**20%**), through GitHub classroom, Due: Aug 26, 11:55pm. This is an individual assessment.
- Assignment 3 (**20%**): through GitHub classroom, Due: Sep 16, 11:55pm. This is an individual assessment.
- Assignment 4, parts 1 and 2 (**20%** each), through GitHub classroom, Due: Oct 7, 11:55pm and Oct 28, 11:55pm.

GitHub Classroom

We are going to use GitHub Classroom (etc5521 2024: Diving Deeper into Data Exploration) to distribute assignment templates and keep track of your assignment progress.

1. Clone the first assignment by clicking on the link given in Moodle.
2. Once you have accepted it, you will get a cloned copy on your own GitHub account. It is a private repo, which means you and the teaching staff will be the only people with access.
3. If you need some help getting started, check this information.
4. The week 1 tutorial is the best place to get help with GitHub.

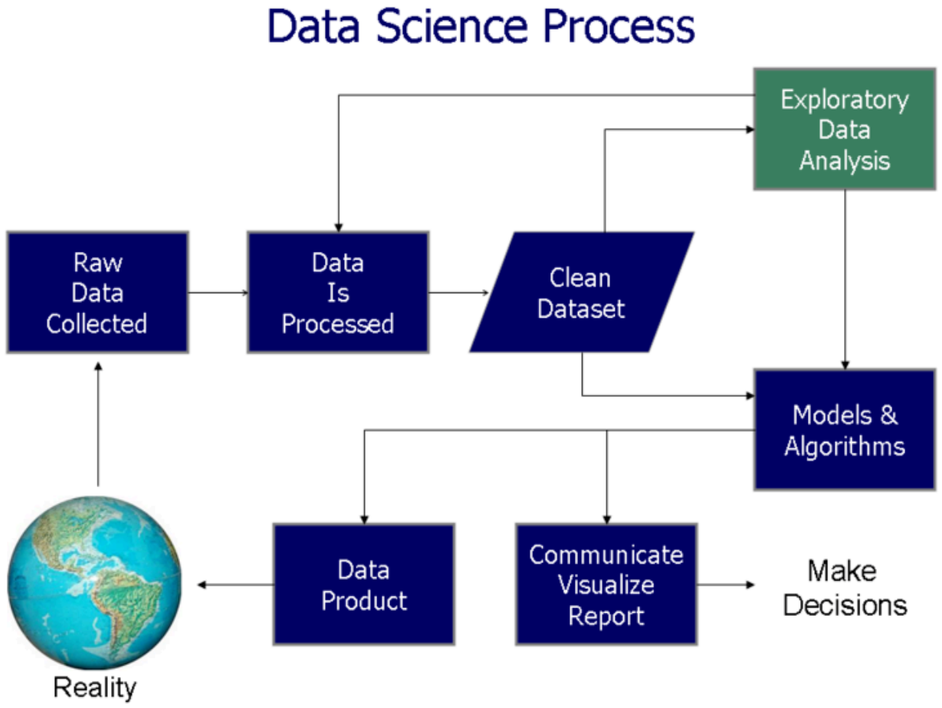
Why exploratory data analysis?

What's special about exploring data, in contrast to confirmatory data analysis?

Let's look at some common definitions and quotes of "exploratory data analysis".

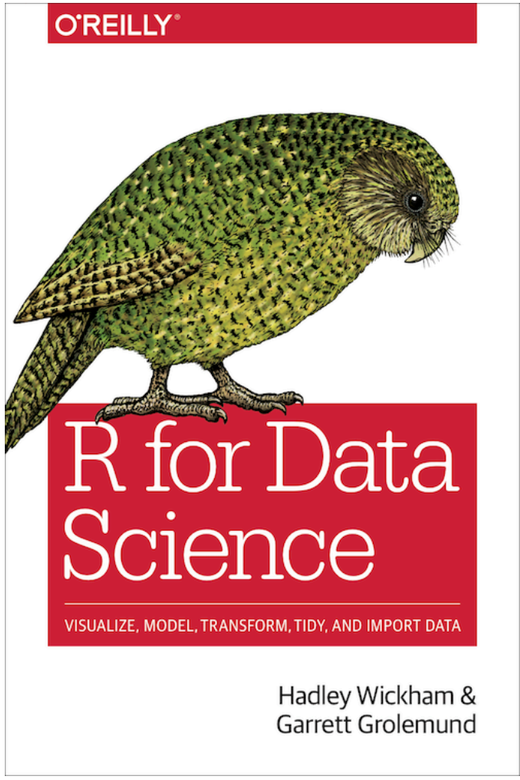
In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

https://en.wikipedia.org/wiki/Exploratory_data_analysis



> EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends.

https://r4ds.had.co.nz/exploratory-data-analysis.html



> Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to (1) maximize insight into a data set; (2) uncover underlying structure; (3) extract important variables; (4) detect outliers and anomalies; (5) test underlying assumptions; (6) develop parsimonious models; and (7) determine optimal factor settings.

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

What is Exploratory Data Analysis (EDA)? (1) How to ensure you are ready to use machine learning algorithms in a project? (2) How to choose the most suitable algorithms for your data set? (3) How to define the feature variables that can potentially be used for machine learning?

https://www.kaggle.com/pavansanagapati/a-simple-tutorial-on-exploratory-data-analysis

> EDA is necessary for the next stage of data research. If there was an analogy to exploratory data analysis, it would be that of a painter examining their tools and available time, before deciding on what best to paint.

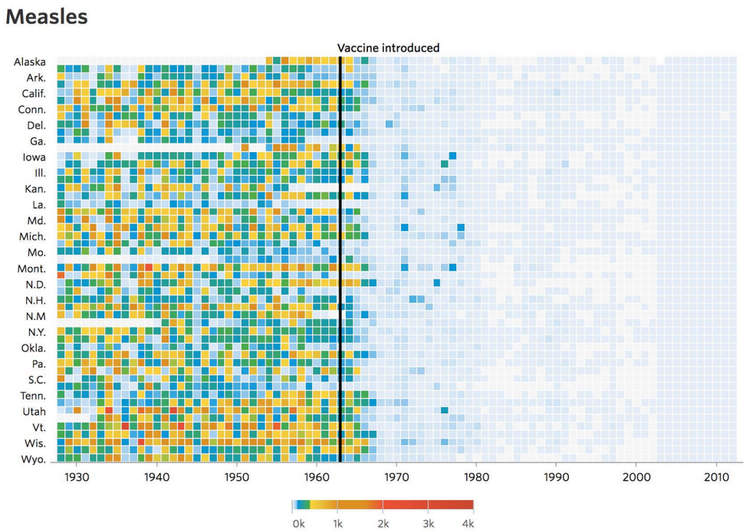
https://seleritysas.com/blog/2020/05/08/exploratory-data-analysis-and-its-role-in-improving-business-operations/

> These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data.

https://www.coursera.org/learn/exploratory-data-analysis#syllabus

> The purpose of doing the Exploratory Data Analysis or EDA is to find new information in data. The understanding of EDA that practitioners may not aware of, is the EDA uses a visually-examined dataset to understand and summarize the main characteristics of the dataset without having a prior hypothesis or relying upon statistical models.

https://towardsdatascience.com/if-you-dont-find-anything-new-you-don-t-do-eda-right-d356f9995098



None of these capture what this course is about

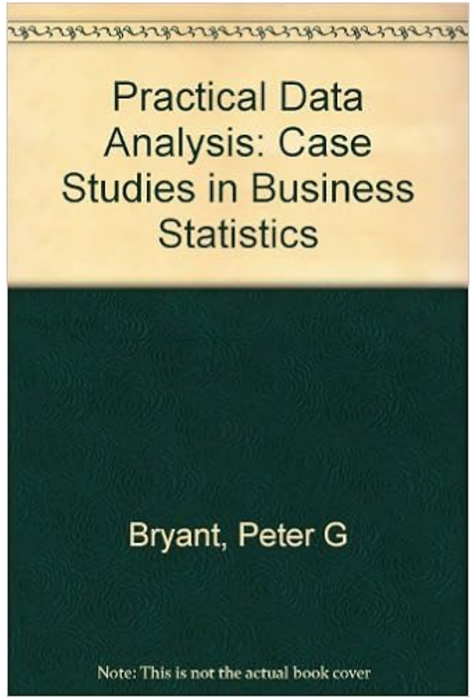
https://www.gocomics.com/calvinandhobbes/2015/08/26

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.




```
1 library(tidyverse)
2 tips <- read_csv("http://ggobi.org/book/data/t
```

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

What is tipping?

- When you're dining at a full-service restaurant
  - Tip 20 percent of your full bill.
- When you grab a cup of coffee
  - Round up or add a dollar if you're a regular or ordered a complicated drink.
- When you have lunch at a food truck
  - Drop a few dollars into the tip jar, but a little less than you would at a dine-in spot.
- When you use a gift card
  - Tip on the total value of the meal, not just what you paid out of pocket.

The basic rules of tipping that everyone should know about

Recommended procedure in the book

- *Step 1*: Develop a model
  - Should the response be `tip` alone and use the total bill as a predictor?
  - Should you create a new variable `tip rate` and use this as the repsonse?
- *Step 2*: Fit the full model with sex, smoker, day, time and size as predictors
- *Step 3*: Refine model: Should some variables should be dropped?
- *Step 4*: Check distribution of residuals
- *Step 5*: Summarise the model, if X=something, what would be the expected tip

Step 1

Calculate tip % as tip/total bill × 100

```
1 tips <- tips %>%
2   mutate(tip_pct = tip/totbill * 100)
```

Note: Creating new variables (sometimes called feature engineering), is a common step in any data analysis.

Step 2 Fit

Fit the full model with all variables

```
1 tips_lm <- tips %>%
2   select(tip_pct, sex, smoker, day, time, size) %>%
3   lm(tip_pct ~ ., data=.)
```

Step 2 Model summary

```
library(broom)
library(kableExtra)
tidy(tips_lm) %>%
  kable(digits=2) %>%
  kable_styling()
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.66	2.49	8.29	0.00
sexM	-0.85	0.83	-1.02	0.31
smokerYes	0.36	0.85	0.43	0.67
daySat	-0.18	1.83	-0.10	0.92
daySun	1.67	1.90	0.88	0.38
dayThu	-1.82	2.32	-0.78	0.43
timeNight	-2.34	2.61	-0.89	0.37
size	-0.96	0.42	-2.28	0.02

```
glance(tips_lm) %>%
  select(r.squared, statistic,
         p.value) %>%
  kable(digits=3)
```

r.squared	statistic	p.value
0.042	1.5	0.17

🤔 Which variable(s) would be considered important for predicting tip %?

Step 3: Refine model

```
tips_lm <- tips %>%
  select(tip_pct, size) %>%
  lm(tip_pct ~ ., data=.)
tidy(tips_lm) %>%
  kable(digits=2) %>%
  kable_styling()
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.44	1.12	16.5	0.00
size	-0.92	0.41	-2.2	0.03

```
glance(tips_lm) %>%
  select(r.squared, statistic, p.value) %>%
  kable(digits=3)
```

r.squared	statistic	p.value
0.02	5	0.026

Model summary

$$\widehat{\text{tip}} = 18.44 - 0.92 \times \text{size}$$

As the size of the dining party increases by one person the tip decreases by approximately 1%.

Model assessment

$R^2 = 0.02.$

This dropped by half from the full model, even though no other variables contributed significantly to the model. It might be a good step to examine interaction terms.

What does $R^2 = 0.02$ mean?

Model assessment

$R^2 = 0.02$ means that size explains just 2% of the variance in tip %. This is a very weak model.

And $R^2 = 0.04$ is also a very weak model.

What do the F statistic and p-value mean?

What do the t statistics and p-value associated with model coefficients mean?

Overall model significance

Assume that we have a random sample from a population. Assume that the model for the population is

$$\widehat{\text{tip}} = \beta_0 + \beta_1 \text{sexM} + \ldots + \beta_7 \text{size}$$

and we have observed

$$\widehat{\text{tip}} = b_0 + b_1 \text{sexM} + \ldots + b_7 \text{size}$$

The F statistic refers to

$$H_o : \beta_1 = \ldots = \beta_7 = 0 \text{ vs } H_a : \text{at least one is not } 0$$

The p-value is the probability that we observe the given F value or larger, computed assuming H_o is true.

Term significance

Assume that we have a random sample from a population. Assume that the model for the population is

$$\widehat{\text{tip}} = \beta_0 + \beta_1 \text{sexM} + \ldots + \beta_7 \text{size}$$

and we have observed

$$\widehat{\text{tip}} = b_0 + b_1 \text{sexM} + \ldots + b_7 \text{size}$$

The t statistics in the coefficient summary refer to

$$H_o : \beta_k = 0 \quad \text{vs} \quad H_a : \beta_k \neq 0$$

The p-value is the probability that we observe the given t value or more extreme, computed assuming H_o is true.

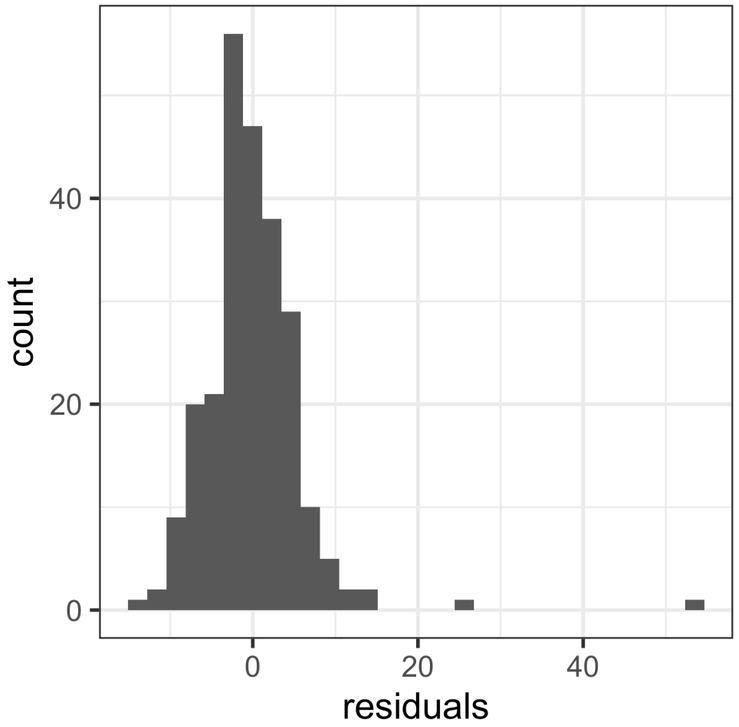
Model diagnostics (MD)

Normally, the final model summary would be accompanied diagnostic plots

- observed vs fitted values to check strength and appropriateness of the fit
- univariate plot, and normal probability plot, of residuals to check for normality
- in the simple final model like this, the observed vs predictor, with model overlaid would be advised to assess the model relative to the variability around the model
- when the final model has more terms, using a partial dependence plot to check the relative relationship between the response and predictors would be recommended.

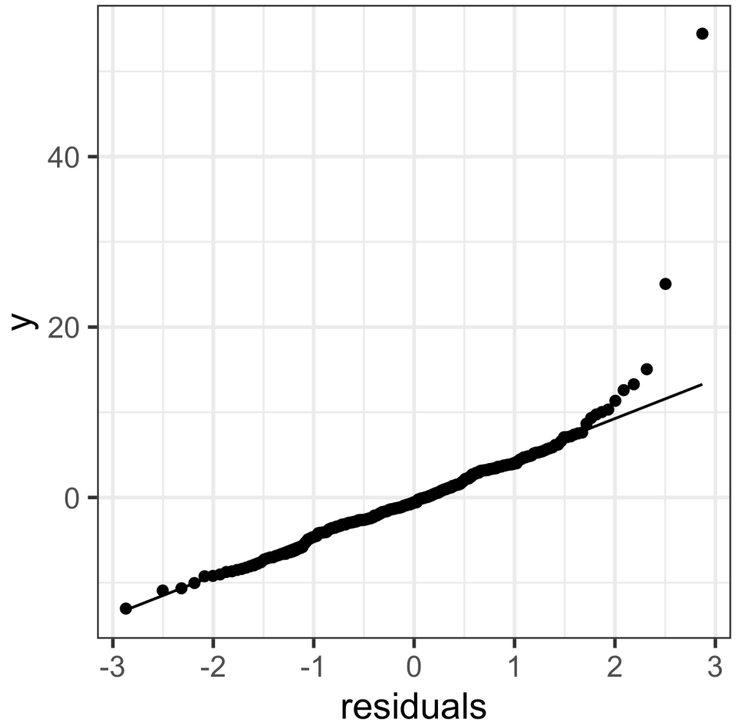
Residual plots

```
1 tips_aug <- augment(tips_lm)
2 ggplot(tips_aug,
3        aes(x=.resid)) +
4   geom_histogram() +
5   xlab("residuals")
```



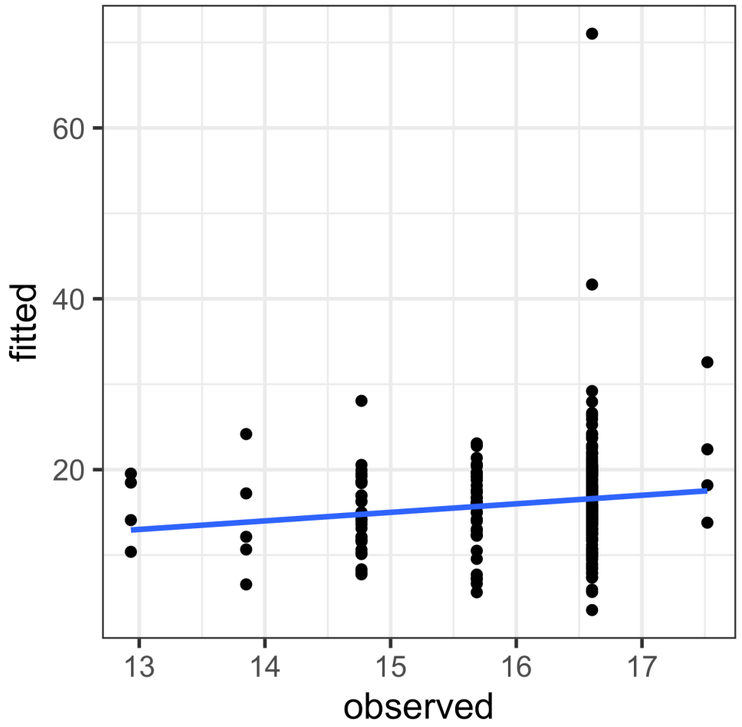
Residual normal probability plots

```
ggplot(tips_aug,
       aes(sample=.resid)) +
  stat_qq() +
  stat_qq_line() +
  xlab("residuals") +
  theme(aspect.ratio=1)
```



Fitted vs observed

```
ggplot(tips_aug,
       aes(x=.fitted, y=tip_pct)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  xlab("observed") +
  ylab("fitted")
```

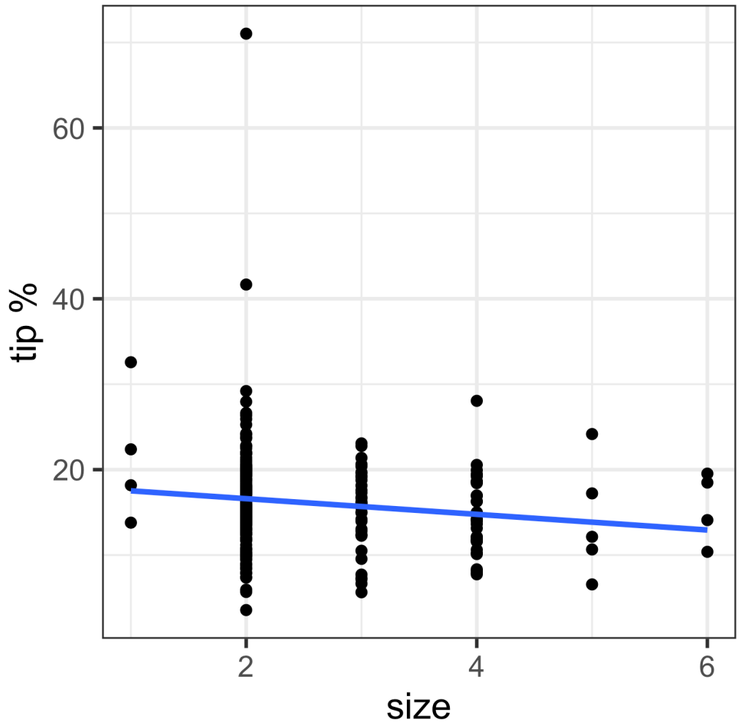


“Model-in-the-data-space”

```
ggplot(tips_aug,
       aes(x=size, y=tip_pct)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ylab("tip %")
```

> The fitted model is overlaid on a plot of the data. This is called “model-in-the-data-space” (Wickham et al, 2015).

All the plots on the previous three slides: histogram of residuals, normal probability plot, fitted vs residuals are considered to be “data-in-the-model-space”. *Stay tuned for more discussion on this later.*



The result of this work might leave us with

a model that could be used to impose a dining/tipping policy in restaurants (see here)

but it should also leave us with an unease that this policy is based on weak support.


BANDAR
(619)238-0101

10:28:33PM [redacted]at
Chk#00066 Tbl: A3 Srv: Eduardo M.

** VISA **

Cardholder: BOLOxxxxxxxxxxxxxxx
Acct#4547xxxxxxxx1188 Exp: 02/09
Seq#57 Appr#066639

Purchases: 53.00

Purchase amount already includes:
18% GRATUITY

Additional Gratuity: _____

TOTAL: $3,00

CUSTOMER COPY

This receipt is for your records.
Enter the tip and total amounts from
the merchant copy onto this receipt.

Summary

> Plots as we have just seen, associated with pursuit of an answer to a specific question may be best grouped into the category of “model diagnostics (MD)”.

There are additional categories of plots for data analysis that include initial data analysis (IDA), descriptive statistics. Stay tuned for more on these.

A separate and big area for plots of data is for communication, where we already know what is in the data and we want to communicate the information as best possible.

> When exploring data, we are using data plots to discover things we **didn’t already know**.

What did this analysis miss?

General strategy for EXPLORING DATA

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

- 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?

```
1 glimpse(tips)
```

```
Rows: 244
Columns: 9
$ obs      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1…
$ totbill <dbl> 17.0, 10.3, 21.0, 23.7, 24.6, 25…
$ tip      <dbl> 1.0, 1.7, 3.5, 3.3, 3.6, 4.7, 2.…
$ sex      <chr> "F", "M", "M", "M", "F", "M", "M…
$ smoker   <chr> "No", "No", "No", "No", "No", "N…
$ day      <chr> "Sun", "Sun", "Sun", "Sun", "Sun…
$ time     <chr> "Night", "Night", "Night", "Nigh…
$ size     <dbl> 2, 3, 3, 2, 4, 4, 2, 4, 2, 2, 2,…
$ tip_pct <dbl> 5.9, 16.1, 16.7, 14.0, 14.7, 18.…
```

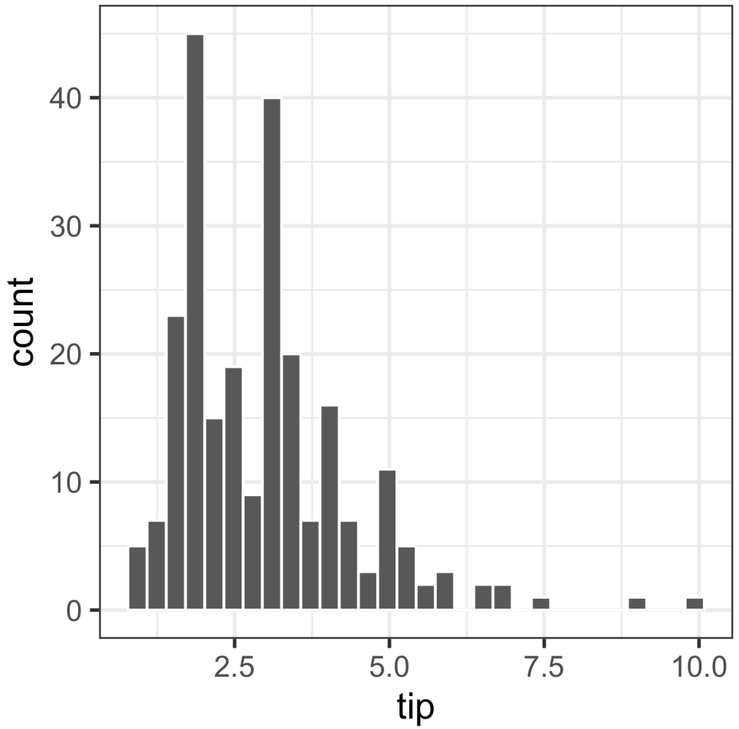
Look at the distribution of quantitative variables tips, total bill.

Examine the distributions across categorical variables.

Examine quantitative variables relative to categorical variables

Distributions of tips

```
1 ggplot(tips,
2        aes(x=tip)) +
3   geom_histogram(
4        colour="white")
```

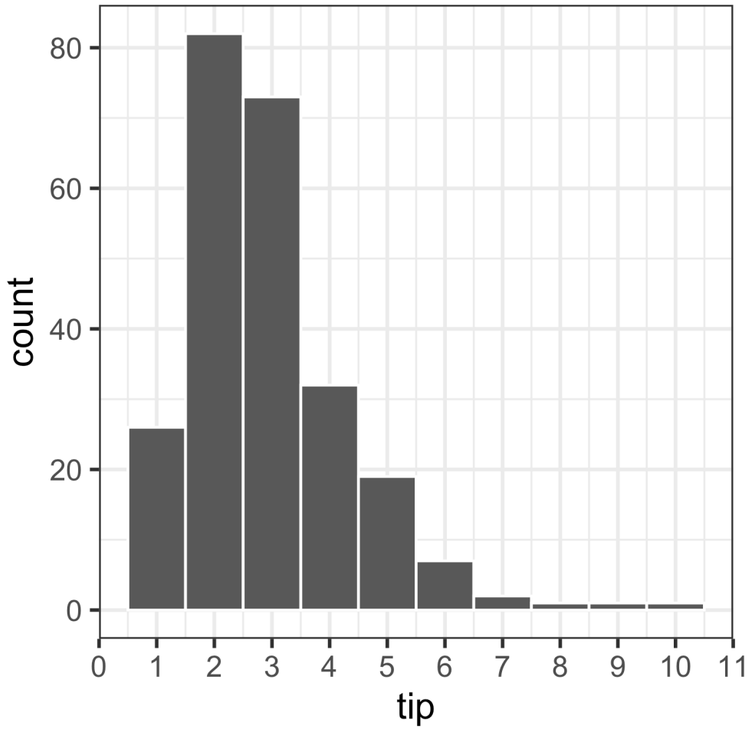



Because, one binwidth is never enough ...

Distributions of tips

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

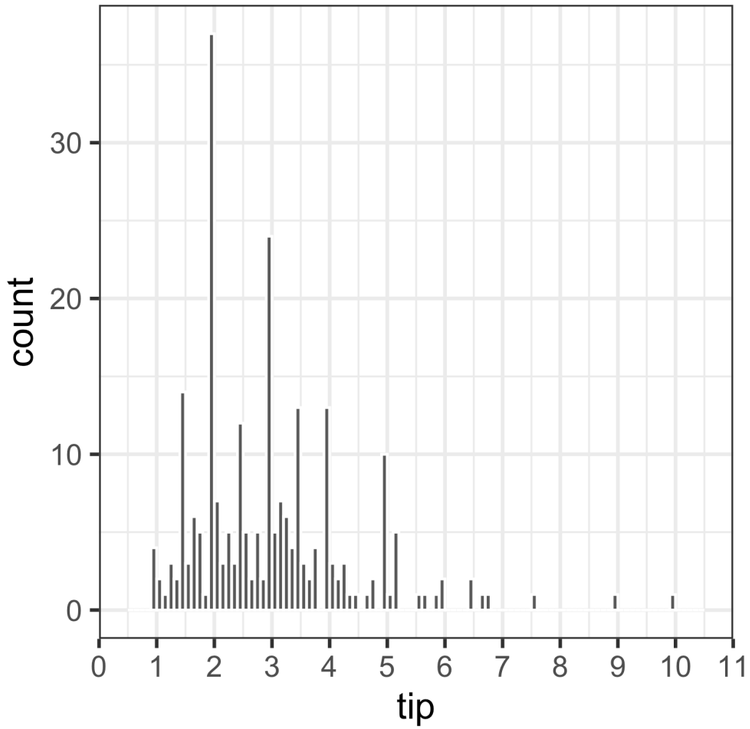
Big fat bins. Tips are skewed, which means most tips are relatively small.



Distributions of tips

```
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

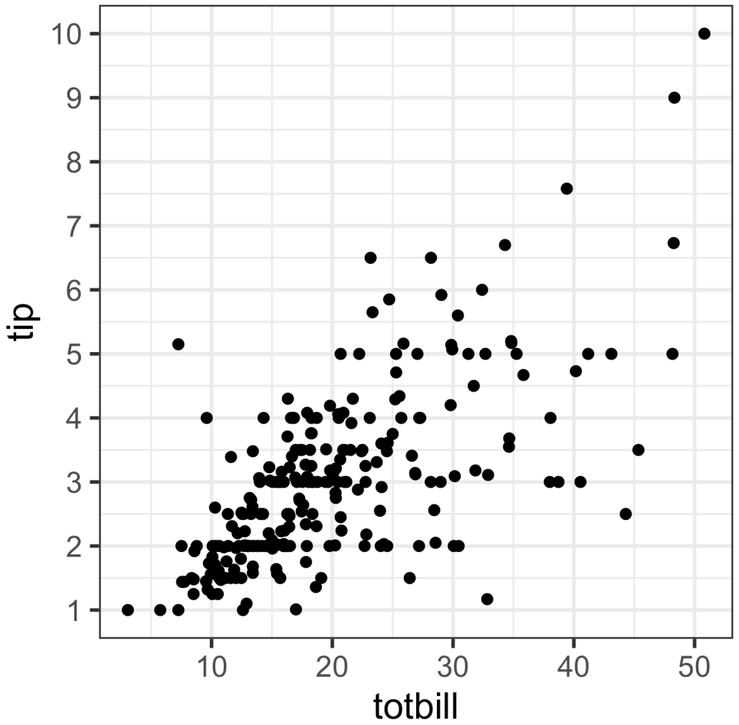
but I've already done this, and we don't learn anything more about the multiple peaks than waht is learned by plotting tips.

Relationship between tip and total

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

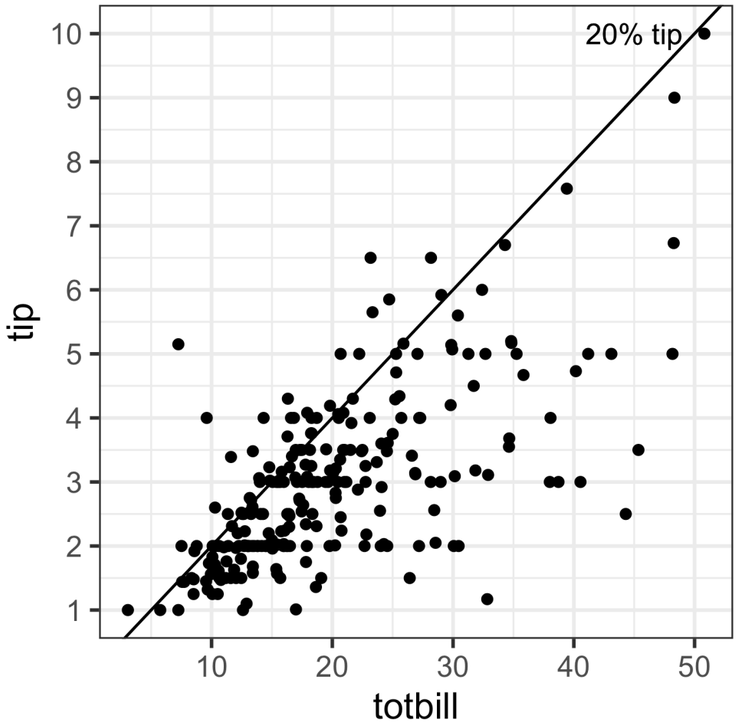
Why is total on the x axis?

Should we add a guideline?



Add a guideline indicating common practice

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



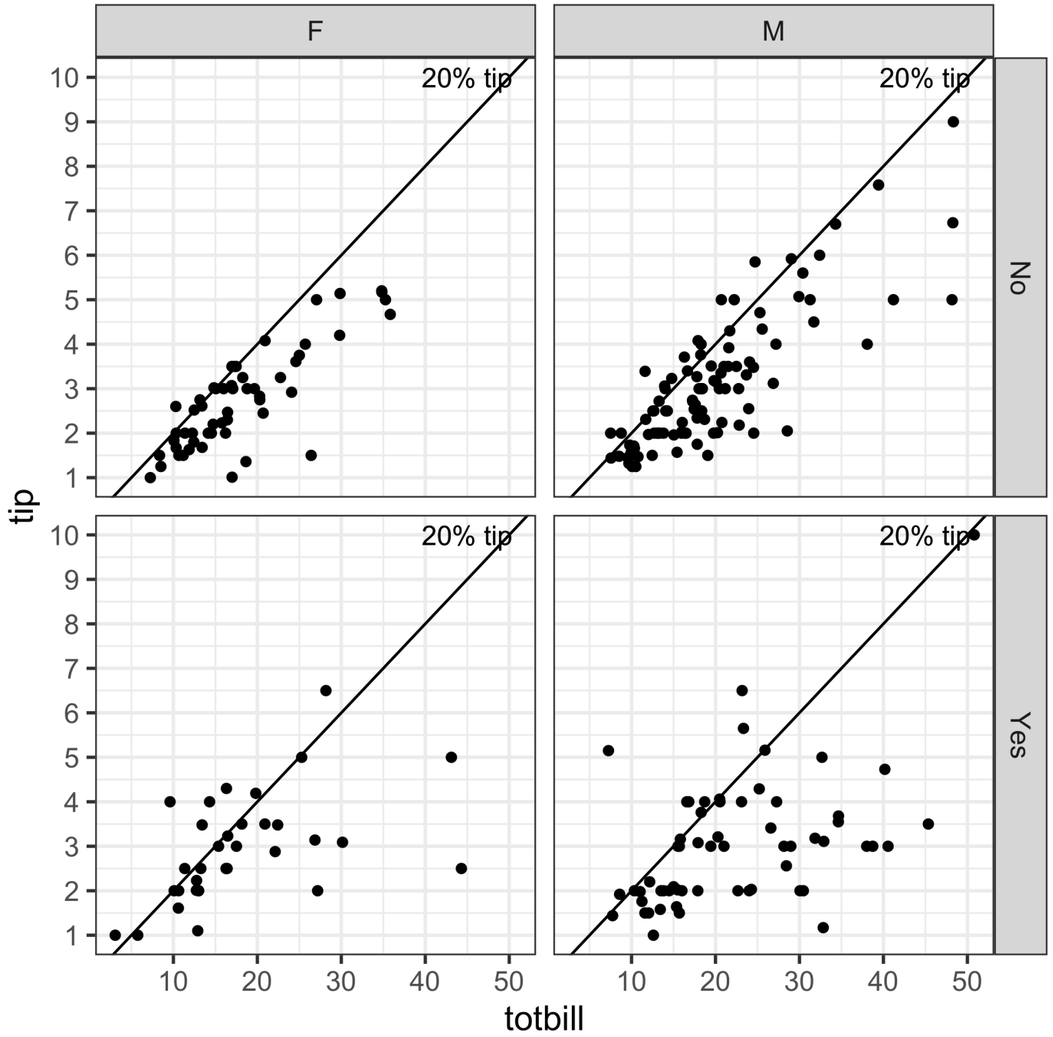
- Most tips less than 20%: Skin flints vs generous diners
- A couple of big tips
- Banding horizontally is the rounding seen previously

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.

Relative to categorical variables

```
1  p + facet_grid(smoker~sex) #<<
```



- 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 insights were missed from the analysis that focused solely on the primary question.

What can go wrong?

How was data collected?

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.

How much can you infer about tipping more broadly?

- Tip has a weak but significant relationship with total bill?
- Tips have a skewed distribution? (More small tips and fewer large tips?)
- Tips tend to be made in nice round numbers.
- People generally under-tip?
- Smokers are less reliable tippers.

Ways to verify, support or refute generalisations

- external information
- other studies/samples
- good choice of calculations and plots
- all the permutations and subsets of measured variables
- computational re-sampling methods (we'll see these soon)

> Poor data collection methods affects every analysis, including statistical or computational modeling.

For this waiter and the restaurant manager, there is some useful information. Like what?

- Service fee for smokers to ensure consistency?
- Assign waiter to variety of party sizes and composition.
- Shifts on different days or time of day (not shown).

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*.

Where do we go from here?

- Methods for single, bivariate, multivariate
  - numerical variables
  - categorical variables
- Methods to accommodate temporal and spatial context
- How to make effective comparisons
- Utilising computational methods to assess what you see is “real”

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