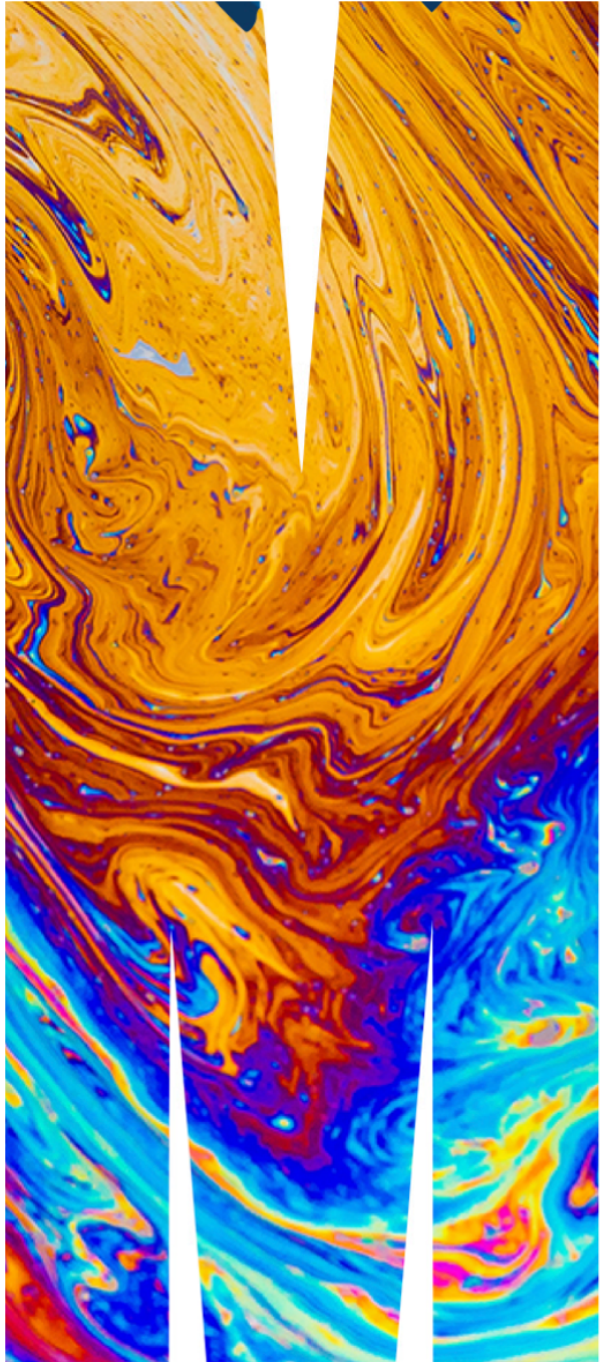
MONASH University

# ETC5521: Exploratory Data Analysis

## Introduction



Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

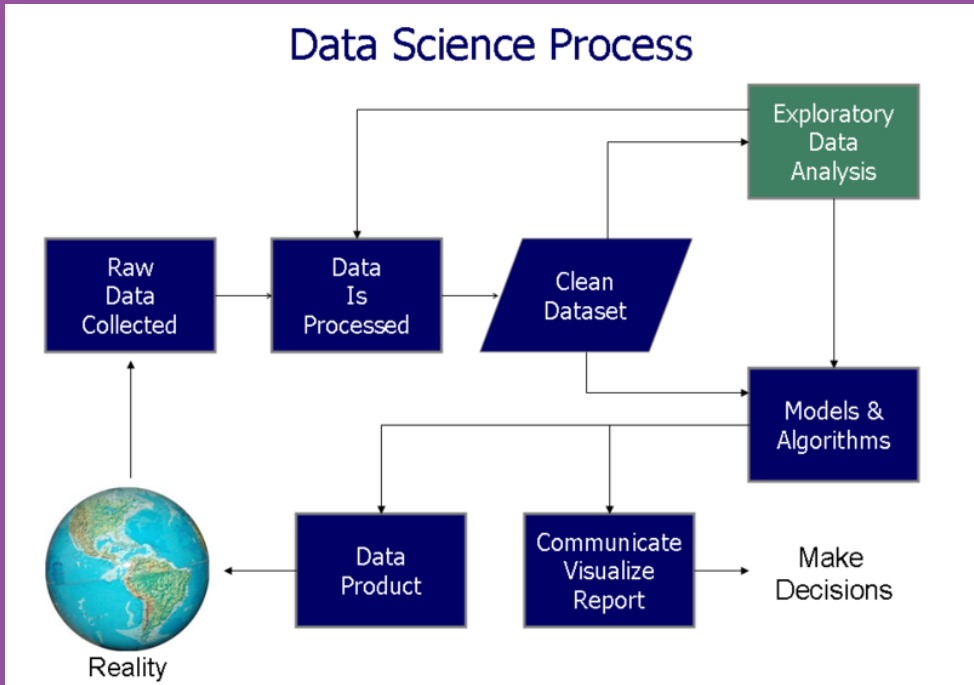
Week 1 - Session 1

# Why this course?

What's special about exploratory data analysis, and different from traditional data analysis?

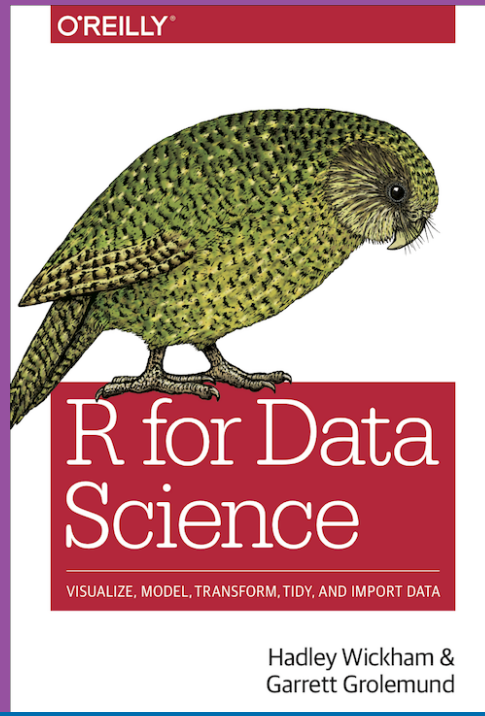
Let's look at some common definitions and quotes

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.



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.

Parsimonious models are simple models with great explanatory predictive power.

They explain data with a minimum number of parameters, or predictor variables. The idea behind parsimonious models stems from Occam's razor, or “the law of briefness” (sometimes called lex parsimoniae in Latin).

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?

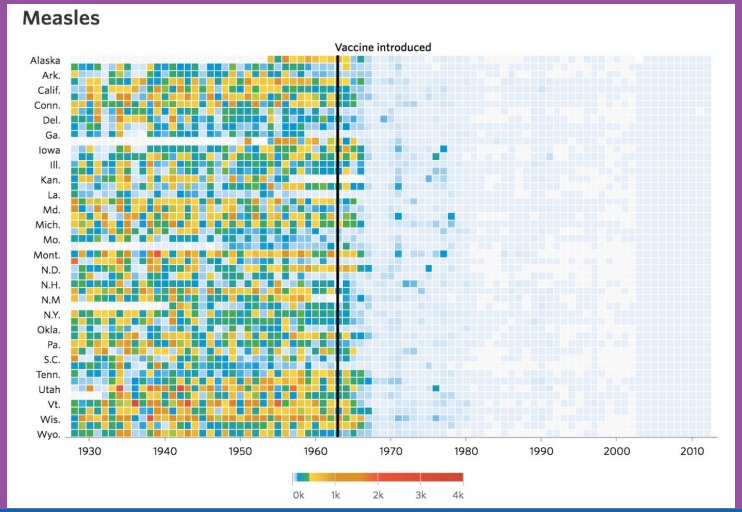

https://www.kaggle.com/pavansanagapati/a-simple-tutorial-on-exploratory-data-analysis

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.

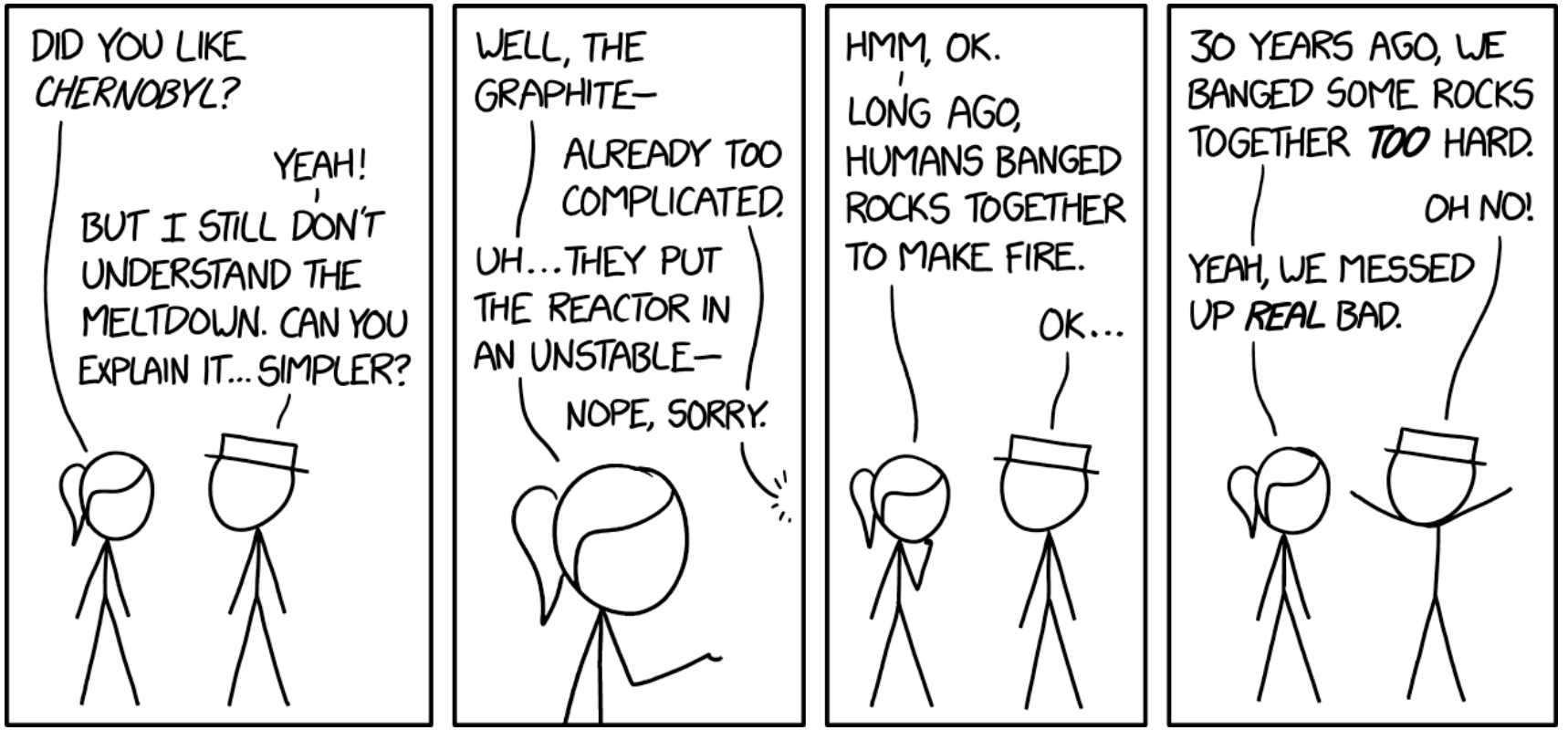
https://seleritysas.com/blog/2020/05/08/exploratory-data-analysis-and-its-role-in-improving-business-operations/

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.

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.

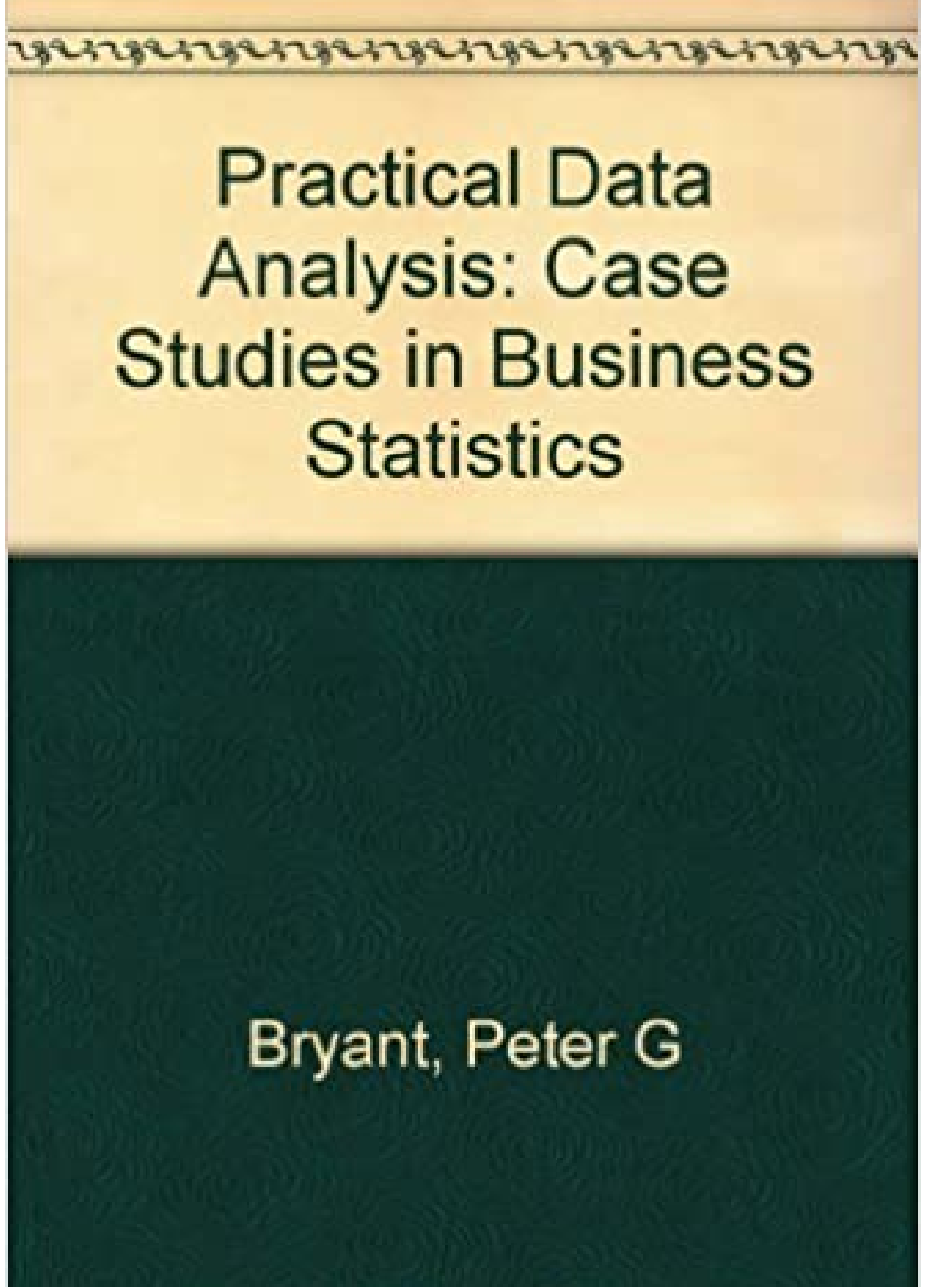



https://towardsdatascience.com/if-you-dont-find-anything-new-you-don-t-do-eda-right-d356f9995098




https://imgs.xkcd.com/comics/chernobyl_2x.png

# 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.

# 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.

# 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

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

# Step 2 Fit

Fit the full model with all variables

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

# Step 2 Model summary

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

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

| 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 |

| r.squared | statistic | p.value |
|---|---|---|
| 0.042 | 1.479 | 0.175 |

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

🔧 *Complete the zoom poll*

# Step 3: Refine model

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

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

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 18.44 | 1.12 | 16.47 | 0.00 |
| size | -0.92 | 0.41 | -2.25 | 0.03 |

| r.squared | statistic | p.value |
|---|---|---|
| 0.02 | 5.042 | 0.026 |

# Model summary

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

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?

$R^2 = 0.02$ 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 coeficients mean?

# Overall model significance

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

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

and we have observed

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

The $F$ statistic refers to

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

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

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

and we have observed

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

The $t$ statistics in the coefficient summary refer to

$$H_o : \beta_k = 0 \;\; vs \;\; H_a : \beta_k \neq 0$$

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

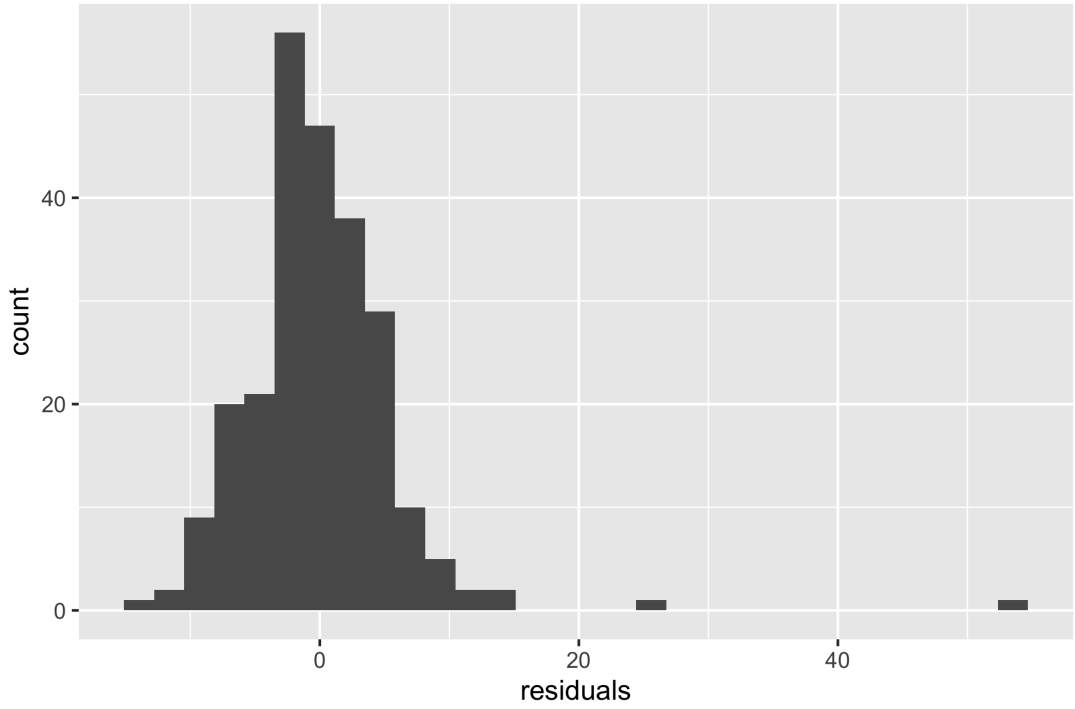
univariate plot, and normal probability plot, of residuals to check for normality

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.

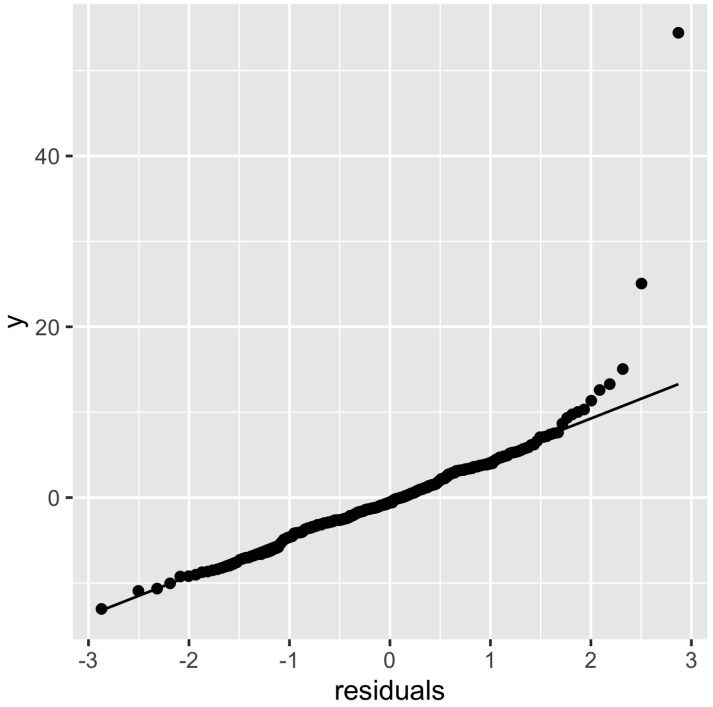
# Residual plots

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



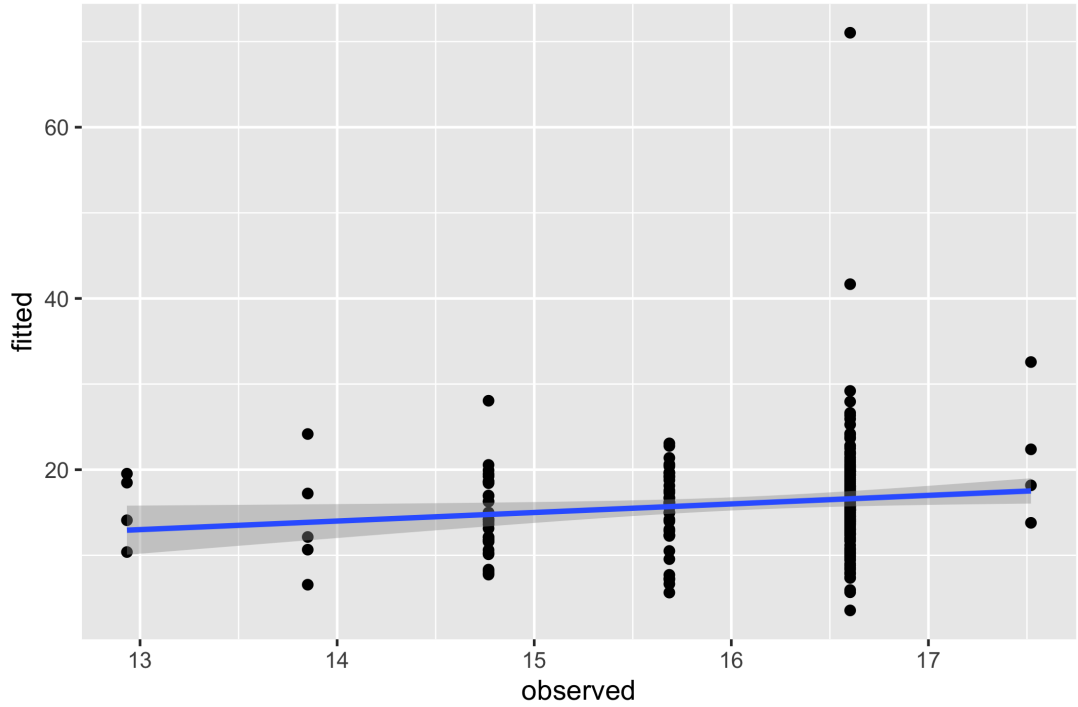
# Residual normal probability plots

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



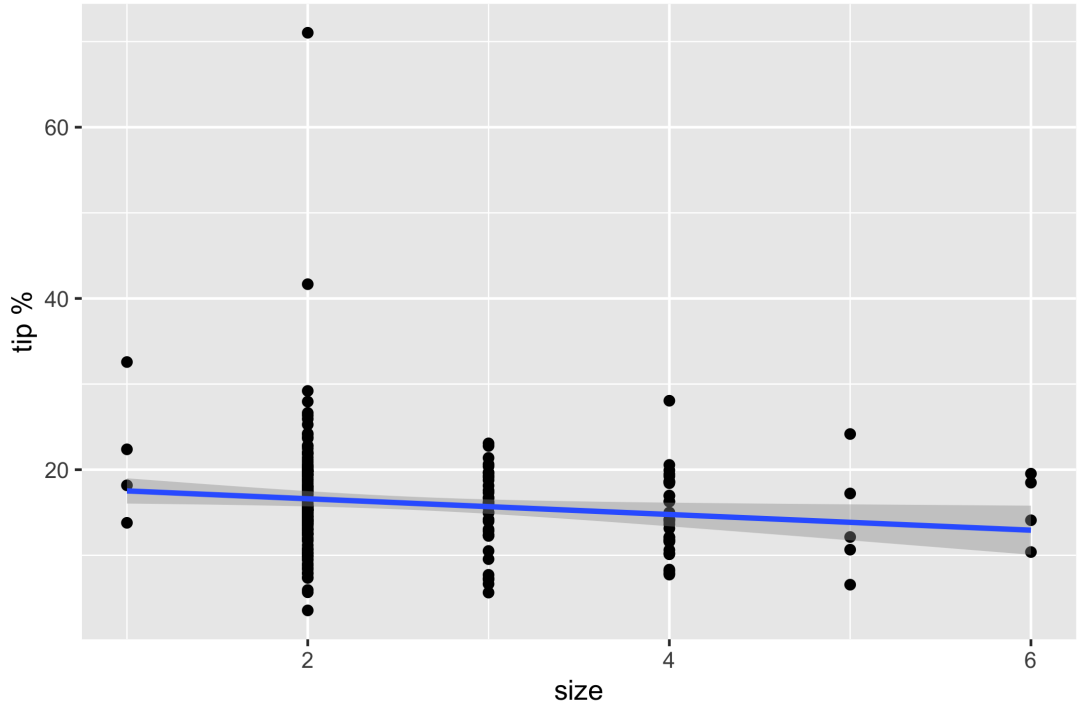
# Fitted vs observed

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



# Model in the data space

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



# The result of this work would leave us with

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

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


BANDAR
(619)238-0101

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

** VISA **

Cardholder: BOLOxxxxxxxxxxxxxx
Acct#4547xxxxxxx1188 Exp: 02/09
Seq#57 Appr#066639

Purchases: 53.00

Purchase amount already includes:
18% GRATUITY

Additional Gratuity: ____

TOTAL: 53,00

CUSTOMER COPY

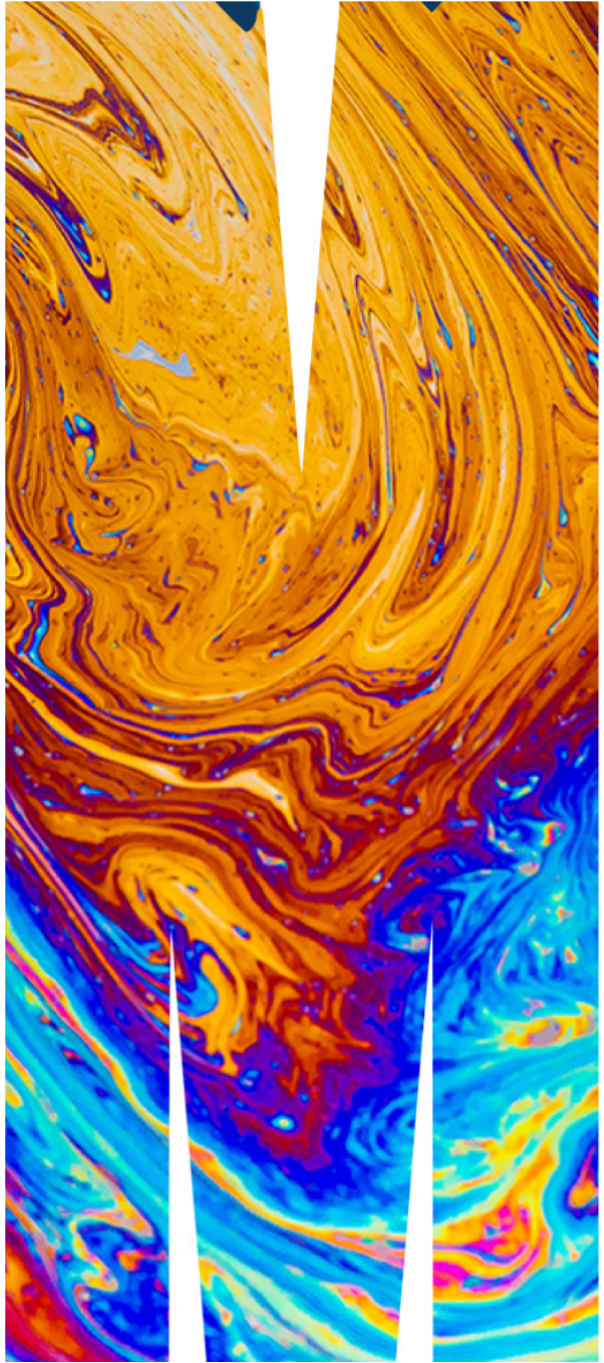
This receipt is for your records.
Enter the tip and total amounts from
the merchant copy onto this receipt.

**Plots as we have just seen, associated with pursuit of an answer to a specific question may be best grouped into the category of "initial data analysis (IDA)" or "model diagnostics (MD)".**

Stay tuned for more on this area later.

MONASH University





Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 1 - Session 1