

Exploratory Data Analysis using R

Dr. Kapil Kumar

Associate Professor, Department of Statistics
Chaudhary Charan Singh University, Meerut, India
kapilstats@gmail.com

UGC-MMTTC Short Term Programme on Applications of Computational Software
in Mathematical Sciences and Engineering (19–24 May 2025)

UGC-Malaviya Mission Teacher Training Centre, Central University of Haryana,
Mahendergarh

22 May 2025 · Day 4 lecture

Plan of the session

1. The EDA mindset, tools and data
2. Univariate EDA: one variable at a time
3. Bivariate EDA: pairs of variables
4. Multivariate EDA: many variables at once
5. An EDA checklist and take-aways

Case studies: `mtcars` (cars), `iris` (flowers) and the *Titanic* passengers.

By the end you can summarise and plot a single variable, detect outliers and missing values, choose the right plot and test for a pair of variables, spot a lurking variable, and explore many variables with a pairs plot, a correlation heatmap, PCA and clustering.

Where we left off yesterday

Session 1, *An Introduction to R*, gave us the tools we now assume:

- Scripts in an RStudio project; `read.csv()` with relative paths
- Vectors, matrices, lists, data frames and factors; indexing with `[ ]`, `$` and `subset()`
- Base graphics: `plot()`, `hist()`, `boxplot()`, plus low-level additions
- Functions, `if`, `for` and the `apply()` family
- A first *t*-test and a first regression, with residual checks

Today We turn those tools on real data sets and ask what the data can tell us *before* we commit to a model: exploratory data analysis.

Yesterday's lesson With the printer sales we plotted first and only then fitted the line. Today that habit becomes a method.

PART 1

The EDA mindset

What is exploratory data analysis?

- Named and promoted by **John W. Tukey** (1977), who also introduced the boxplot
- Look at the data *before* and *while* modelling: graphs first, numbers alongside
- Aims: see distributions and relationships; find errors, gaps and outliers; check assumptions; **suggest** hypotheses
- Iterative: every plot raises the next question

Tukey “The greatest value of a picture is when it forces us to notice what we never expected to see.”

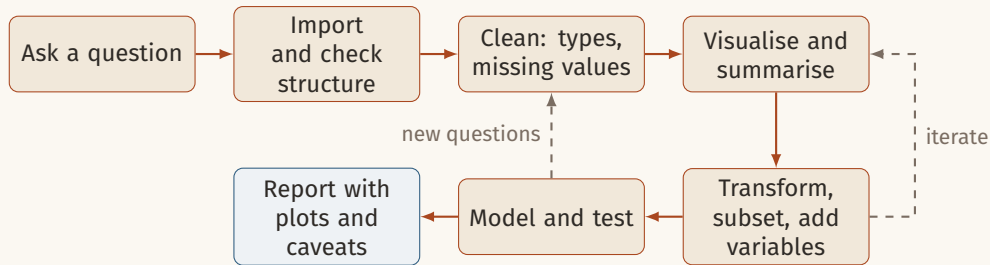
Exploratory vs confirmatory EDA *generates* questions; confirmatory analysis *tests* them, ideally on data that did not suggest them.

A toolbox organised by the question

	Graphical	Numerical	To confirm
Univariate one variable	histogram, density, box-plot, bar chart, Q-Q plot	mean, median, SD, IQR, skewness, kurtosis, counts	normality checks
Bivariate two variables	scatter plot, side-by-side boxplots, filled bar chart	correlation, two-way table, group means	<code>cor.test()</code> , <code>t.test()</code> , <code>aov()</code> , <code>chisq.test()</code>
Multivariate many variables	pairs plot, heatmap, biplot, cluster plot, facets	correlation matrix, principal components, clusters	regression with diagnostics

Rule of thumb Match the tool to the variable *types*, and pair every number with a picture.

The EDA cycle



Most of the time goes into the first four boxes. A model fitted without them can be precise and still wrong.

Packages used in this session

- `ggplot2`, `gridExtra`: graphics, panels
- `dplyr`, `tidyr`: grouped summaries
- `psych`, `moments`: descriptive statistics
- `GGally`, `corrplot`: pairs plot, heatmap
- `factoextra`: PCA and cluster plots
- `lmtest`, `car`: regression diagnostics
- `titanic`: passenger-level Titanic data

```
# install.packages(c("GGally", "corrplot", ...))
library(ggplot2); library(gridExtra)
library(dplyr); library(tidyr)
library(psych); library(moments)
library(GGally); library(corrplot)
library(factoextra)
library(lmtest); library(car)
library(titanic)
col_main <- "#A9471F"; col_alt <- "#2F5D80"
col_species <- c(setosa = "#A9471F",
                  versicolor = "#2F5D80",
                  virginica = "#7C8B2E")
```

Three classic data sets

Data	Size	One row is	Origin
mtcars	32×11	a car model	<i>Motor Trend</i> magazine, 1974: fuel use and design of 1973–74 models
iris	150×5	a flower	Anderson's measurements used by Fisher (1936): 50 flowers of each of three species
titanic_train	891×12	a passenger	titanic package: the training part of the passenger list
Titanic	4-way table	a cell count	base R: all 2,201 people on board, crew included

First look: structure before statistics

```
titanic <- titanic_train; cars <- mtcars # working copies
titanic$Survived <- factor(titanic$Survived, labels = c("No", "Yes"))
titanic$Pclass <- factor(titanic$Pclass, labels = c("1st", "2nd", "3rd"))
cars$am <- factor(cars$am, labels = c("automatic", "manual"))
sapply(titanic[, c("Survived", "Pclass", "Sex", "Age")], class)

##      Survived      Pclass      Sex      Age
##      "factor"    "factor" "character" "numeric"
```

Watch out Codes such as 0/1 are categories: make them labelled factors first.
Empty strings are not NA: Cabin is blank for 687 passengers, Embarked for 2.

PART 2

Univariate EDA

Numerical summary of fuel economy

```
summary(cars$mpg)           # miles per gallon
d <- describe(cars$mpg)     # psych
d[, c("n", "mean", "sd", "median", "mad", "se")]

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.4    19.2    20.1    22.8    33.9
##      n  mean   sd median  mad   se
## X1 32 20.09 6.03   19.2 5.41 1.07
```

Reading it A typical car does about 19.2 mpg (median); the middle half lies between 15.425 and 22.8. The mean (20.09) exceeds the median: a hint of a longer right tail. SD 6.03 mpg.

$$\bar{x} = \frac{1}{n} \sum x_i, \quad s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}, \quad SE = s/\sqrt{n}$$
$$IQR = Q_3 - Q_1, \quad mad = 1.4826 \times median |x_i - \tilde{x}|$$

Shape: skewness and kurtosis

Moments: $m_k = \frac{1}{n} \sum (x_i - \bar{x})^k$

Skewness: $g_1 = m_3/m_2^{3/2}$

Kurtosis: $b_2 = m_4/m_2^2$

Normal: $g_1 = 0$, $b_2 = 3$

Same data, different numbers

`moments` gives b_2 (normal 3); `psych` gives the excess kurtosis $m_4/s^4 - 3$ with the $n - 1$ SD (normal 0). Check the definition.

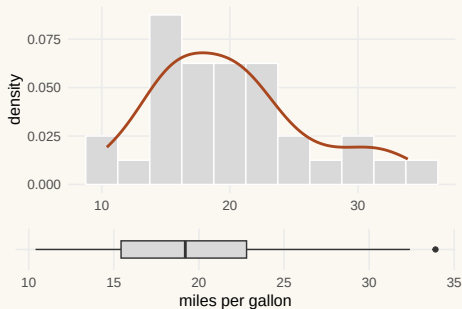
```
skewness(cars$mpg)      # moments: g1
kurtosis(cars$mpg)      # moments: b2
dskurtosis              # psych: m4/s^4 - 3

## [1] 0.6404
## [1] 2.799
## [1] -0.3728
```

Reading it $g_1 = 0.64$: moderate right skew. Excess kurtosis $2.8 - 3 = -0.2$: slightly light tails. Rough, with $n = 32$.

Histogram, density and boxplot

```
p1 <- ggplot(cars, aes(mpg)) +  
  geom_histogram(aes(y = after_stat(density)),  
    binwidth = 2.5, fill = "grey85",  
    colour = "white") +  
  geom_density(colour = col_main, linewidth = 1) +  
  labs(x = NULL, y = "density")  
p2 <- ggplot(cars, aes(mpg, y = "")) +  
  geom_boxplot(fill = "grey85", width = 0.5) +  
  labs(x = "miles per gallon", y = NULL)  
grid.arrange(p1, p2, ncol = 1,  
  heights = c(3, 1.3))
```



Most cars do 15–23 mpg, with a thin right tail of economical cars. Both panels share the x -axis.

Outliers: the $1.5 \times \text{IQR}$ rule

```
Q1 <- quantile(cars$mpg, 0.25)
Q3 <- quantile(cars$mpg, 0.75)
IQR_val <- Q3 - Q1
fence <- c(lower = Q1 - 1.5 * IQR_val,
           upper = Q3 + 1.5 * IQR_val)

fence
cars$mpg[cars$mpg < fence[1] | cars$mpg > fence[2]]
fivenum(cars$mpg)      # Tukey's hinges, as in boxplot()

## lower.25% upper.75%
##      4.363      33.862
## [1] 33.9
## [1] 10.40 15.35 19.20 22.80 33.90
```

Flag x if $x < Q_1 - 1.5 \text{ IQR}$
or $x > Q_3 + 1.5 \text{ IQR}$

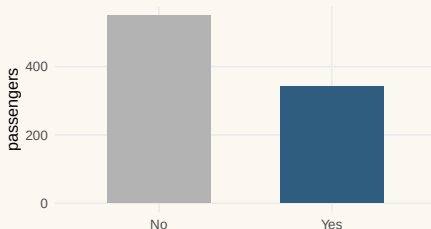
Reading it Only 33.9 mpg (Toyota Corolla) crosses the fence of 33.86. `boxplot()` uses the hinges (upper fence 33.975) and draws no outlier: a borderline but genuine car, not an error.

A categorical variable: who survived?

```
summary(titanic$Survived)           # counts  
round(prop.table(summary(titanic$Survived)), 3)
```

```
##  No  Yes  
## 549 342  
##    No   Yes  
## 0.616 0.384
```

```
ggplot(titanic, aes(Survived)) +  
  geom_bar(width = 0.6,  
           fill = c("grey70", col_alt)) +  
  labs(x = NULL, y = "passengers")
```



Reading it 38.4% of these 891 passengers survived. For a categorical variable, counts and proportions *are* the summary.

Missing values: who has no recorded age?

```
sum(is.na(titanic$Age))
round(100 * mean(is.na(titanic$Age)), 1) # percent
# share of missing ages within each class
round(100 * tapply(is.na(titanic$Age),
                    titanic$Pclass, mean), 1)

## [1] 177
## [1] 19.9
## 1st  2nd  3rd
## 13.9  6.0 27.7
```

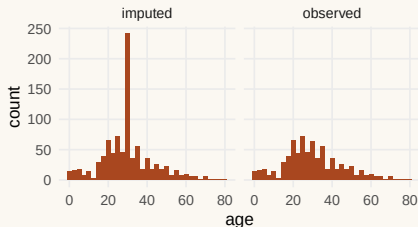
Reading it Age is missing for 177 passengers (19.9%). Missingness depends on class: 27.7% in third class against 6% in second.

So what? The missing ages are not a random sample: dropping or filling them can bias class comparisons. Always ask *why* data are missing.

Mean imputation, and what it costs

```
mean_age <- mean(titanic$Age, na.rm = TRUE)
age_imp <- titanic$Age
age_imp[is.na(age_imp)] <- mean_age
c(sd(titanic$Age, na.rm = TRUE), sd(age_imp))
ages <- data.frame(age = c(titanic$Age, age_imp),
  version = rep(c("observed", "imputed"),
    each = nrow(titanic)))
ggplot(na.omit(ages), aes(age)) +
  geom_histogram(binwidth = 2.5, fill = col_main) +
  facet_wrap(~ version)

## [1] 14.53 13.00
```

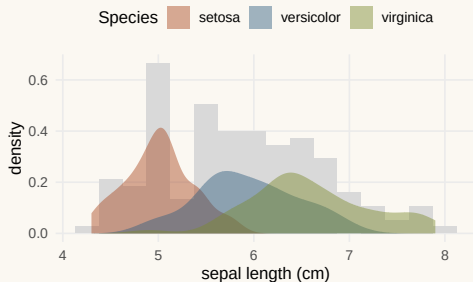


177 copies of 29.7 form a spike; the SD drops from 14.5 to 13.0, so tests look too precise. Better: impute by group, or multiple imputation.

A mixture in disguise: iris sepal length

```
ggplot(iris, aes(Sepal.Length)) +  
  geom_histogram(aes(y = after_stat(density)),  
    binwidth = 0.25, fill = "grey85") +  
  geom_density(aes(fill = Species,  
    y = after_stat(density) / 3),  
    alpha = 0.45, colour = NA) +  
  scale_fill_manual(values = col_species) +  
  labs(x = "sepal length (cm)") +  
  theme(legend.position = "top")
```

Each species' density is scaled by 1/3, its share of the 150 flowers, so the three curves add up to the overall density.



One broad hump, but three species with different centres: a univariate summary can hide groups.

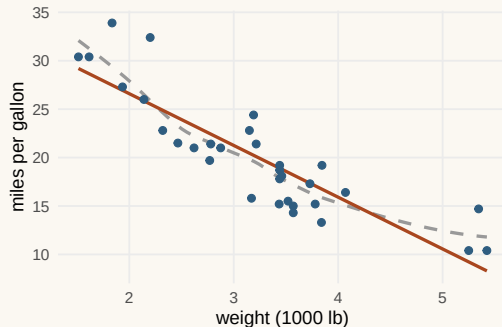
PART 3

Bivariate EDA

Fuel economy and weight

```
ggplot(cars, aes(wt, mpg)) +  
  geom_smooth(method = "loess", se = FALSE,  
             colour = "grey60", linetype = 2) +  
  geom_smooth(method = "lm", se = FALSE,  
             colour = col_main) +  
  geom_point(colour = col_alt, size = 2) +  
  labs(x = "weight (1000 lb)",  
       y = "miles per gallon")
```

Numerical against numerical: a scatter plot, with a straight line (solid) and a flexible smooth (dashed) to compare.



Heavier cars do fewer miles per gallon; the smooth bends a little at the extremes.

Correlation, with a test

```
ct <- cor.test(cars$mpg, cars$wt)
round(c(r = unname(ct$estimate),
        t = unname(ct$statistic),
        p = ct$p.value), 4)
round(ct$conf.int[1:2], 3)      # 95% CI for rho
cor(cars$mpg, cars$wt, method = "spearman")
```

```
##      r      t      p
## -0.8677 -9.5590 0.0000
## [1] -0.934 -0.744
## [1] -0.8864
```

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$H_0 : \rho = 0: \quad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

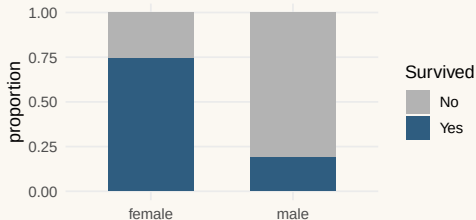
Reading it $r = -0.87$, 95% CI (-0.93, -0.74): a strong negative linear association. Spearman's rank correlation agrees, so a few points do not drive it.

Sex and survival: a two-way table

```
tab_sex <- table(Sex = titanic$Sex,  
                 Survived = titanic$Survived)  
tab_sex  
# row percentages: survival within each sex  
round(100 * prop.table(tab_sex, margin = 1), 1)
```

```
##           Survived  
## Sex           No Yes  
## female  81 233  
## male   468 109  
##           Survived  
## Sex           No Yes  
## female  25.8 74.2  
## male   81.1 18.9
```

```
ggplot(titanic, aes(Sex, fill = Survived)) +  
  geom_bar(position = "fill", width = 0.6) +  
  scale_fill_manual(values = c("grey70",  
                               col_alt)) +  
  labs(x = NULL, y = "proportion")
```



Chi-squared tests of association

```
tab_class <- table(Class = titanic$Pclass,
                   Survived = titanic$Survived)
c1 <- chisq.test(tab_sex)
c2 <- chisq.test(tab_class)
data.frame(X2 = c(c1$statistic, c2$statistic),
           df = c(c1$parameter, c2$parameter),
           p = c(c1$p.value, c2$p.value),
           row.names = c("sex", "class"))

##           X2 df           p
## sex    260.7  1 1.197e-58
## class 102.9  2 4.549e-23
```

$$X^2 = \sum (O - E)^2 / E$$

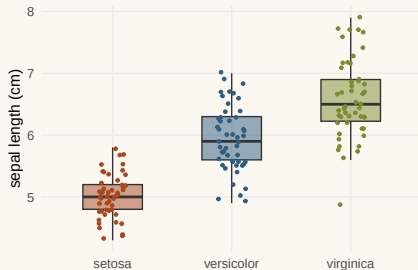
$$E = \text{row total} \times \text{column total} / n$$

$$\text{df} = (r - 1)(c - 1)$$

Reading it Survival depends strongly on sex and on class ($p < 0.001$; all expected counts ≥ 70): 63% in first class, 24% in third. Report such percentages, not only p .

Sepal length by species

```
ggplot(iris, aes(Species, Sepal.Length,  
                 fill = Species)) +  
  geom_boxplot(alpha = 0.5, width = 0.5,  
               outlier.shape = NA) +  
  geom_jitter(aes(colour = Species),  
              width = 0.12, size = 1) +  
  scale_fill_manual(values = col_species) +  
  scale_colour_manual(values = col_species) +  
  labs(x = NULL, y = "sepal length (cm)") +  
  theme(legend.position = "none")
```



Numerical against categorical. Medians rise from setosa to virginica and the spread grows; only one flower (4.9 cm, virginica) lies beyond its whiskers.

One-way ANOVA, and checking its assumptions

```
fit_aov <- aov(Sepal.Length ~ Species, data = iris)
summary(fit_aov)
bartlett.test(Sepal.Length ~ Species, iris)$p.value
oneway.test(Sepal.Length ~ Species, iris)$p.value
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species      2   63.2   31.61    119 <2e-16 ***
## Residuals    147   39.0    0.27
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] 0.0003345
## [1] 1.505e-28
```

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$F = MS_{\text{between}} / MS_{\text{within}}$$

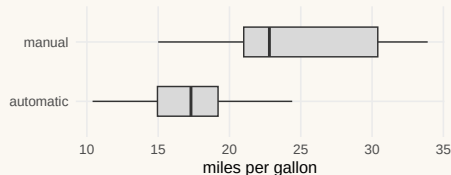
Reading it $F = 119.3$,
 $p < 0.001$: mean sepal length
differs by species. Bartlett's
test rejects equal variances
($p = 0.0003$); Welch's ANOVA
(last line) does not need them
and agrees.

Fuel economy by transmission

```
tt <- t.test(mpg ~ am, data = cars)      # Welch
tt$estimate
round(c(t = unname(tt$statistic), p = tt$p.value,
        ci = tt$conf.int), 3) # CI: auto - manual

## mean in group automatic    mean in group manual
##           17.15              24.39
##      t          p       ci1       ci2
## -3.767    0.001 -11.280    -3.210
```

```
ggplot(cars, aes(mpg, am)) +
  geom_boxplot(fill = "grey85", width = 0.5) +
  labs(x = "miles per gallon", y = NULL)
```



Reading it Manual cars average 7.2 mpg more (95% CI 3.2 to 11.3; $p = 0.0014$). But is the gearbox the reason?

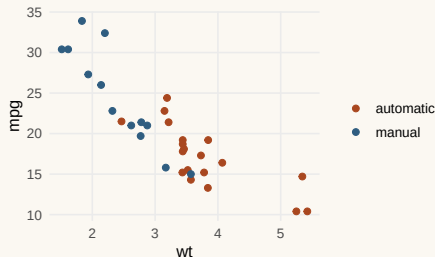
A lurking variable: weight

```
tapply(cars$wt, cars$am, mean) # 1000 lb
fit_wa <- lm(mpg ~ wt + am, data = cars)
coef(summary(fit_wa))["ammanual", c(1, 4)]

## automatic    manual
##      3.769      2.411
## Estimate Pr(>|t|)
## -0.02362  0.98791
```

Reading it Manual cars are lighter; at equal weight the gap is -0.02 mpg ($p = 0.99$).

```
ggplot(cars, aes(wt, mpg, colour = am)) +
  geom_point(size = 2) + labs(colour = NULL) +
  scale_colour_manual(values = c(col_main,
                                col_alt))
```

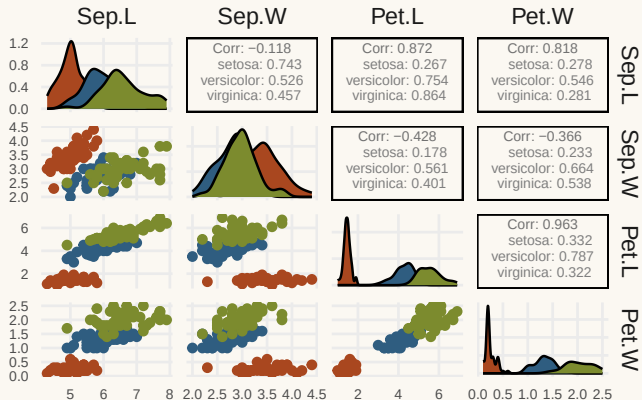


PART 4

Multivariate EDA

Scatterplot matrix of iris

```
ggpairs(iris, columns = 1:4,  
  aes(colour = Species),  
  columnLabels = c("Sep.L", "Sep.W",  
    "Pet.L", "Pet.W"),  
  upper = list(continuous =  
    wrap("cor", size = 2.2,  
      stars = FALSE))) +  
  scale_colour_manual(values =  
    col_species) +  
  scale_fill_manual(values =  
    col_species) +  
  theme(axis.text =  
    element_text(size = 6))
```



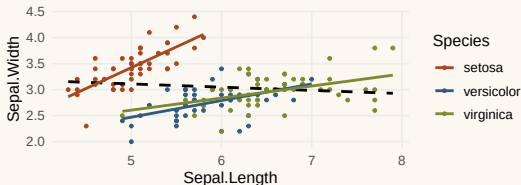
Reading the pairs plot: Simpson's paradox

```
cor(iris$Sepal.Length, iris$Sepal.Width)
apply(split(iris, iris$Species),
      function(d) cor(d$Sepal.Length,
                      d$Sepal.Width))
```

```
## [1] -0.1176
##      setosa versicolor  virginica
##      0.7425      0.5259      0.4572
```

Reading it Pooled: $r = -0.12$, no relation. Within each species: clearly *positive*.

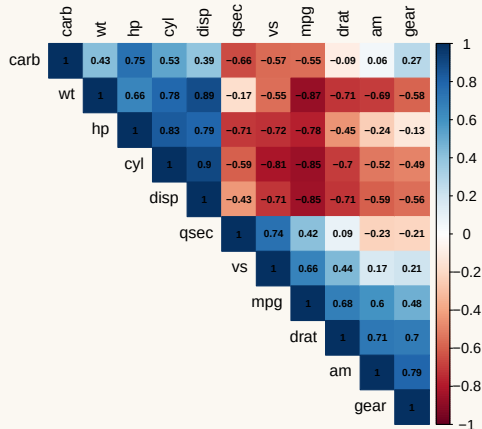
```
ggplot(iris, aes(Sepal.Length, Sepal.Width)) +
  geom_point(aes(colour = Species), size = 1) +
  geom_smooth(method = "lm", se = FALSE,
             colour = "black", linetype = 2) +
  geom_smooth(aes(colour = Species),
             method = "lm", se = FALSE) +
  scale_colour_manual(values = col_species)
```



Correlation heatmap of the car data

```
cor_matrix <- cor(mtcars)
corrplot(cor_matrix, method = "color",
         type = "upper", order = "hclust",
         tl.col = "black", tl.cex = 0.9,
         addCoef.col = "black",
         number.cex = 0.55)
```

Reading it Two blocks, size and power (cyl, disp, hp, wt) against economy (mpg, drat, am, gear); mpg tracks wt most closely ($r = -0.87$).



Principal component analysis

Standardise the p variables;
 R = their correlation matrix.

$$R a_k = \lambda_k a_k, \lambda_1 \geq \dots \geq \lambda_p$$

$$PC_k = a_k^\top z, \text{ variance } \lambda_k, \text{ share } \lambda_k/p$$

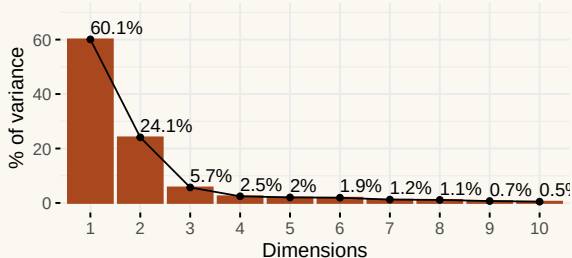
The PCs are uncorrelated and ordered by the variance they carry. Scaling matters: unscaled, disp and hp would dominate through their units. Correlated blocks, as in the heatmap, are what PCA compresses.

```
pca <- prcomp(mtcars, scale. = TRUE)
round(pca$sdev^2, 2)           # eigenvalues
round(summary(pca)$importance[, 1:4], 3)

## [1] 6.61 2.65 0.63 0.27 0.22 0.21 0.14 0.12 0.08
## [10] 0.05 0.02
##
##                               PC1    PC2    PC3    PC4
## Standard deviation          2.571 1.628 0.792 0.519
## Proportion of Variance      0.601 0.241 0.057 0.025
## Cumulative Proportion       0.601 0.842 0.899 0.923
```

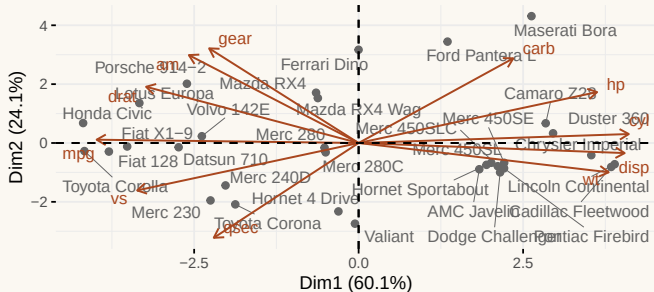
How many components? The scree plot

```
fviz_eig(pca, addlabels = TRUE, ylim = c(0, 70),
         barfill = col_main, barcolor = col_main,
         ggtheme = theme_minimal(base_size = 12)) +
  labs(title = NULL, y = "% of variance")
```



Reading it PC1 carries 60.1% of the total variance and PC2 24.1%: together 84.2%. The curve flattens after two components, and only two eigenvalues exceed 1.

Eleven correlated measurements are well summarised by two new variables.



Reading it **PC1:** size and power (cyl, disp, wt, hp) against economy (mpg). **PC2:** gear, am and carb against quarter-mile time (qsec): sporty versus sedate.

k -means clustering of iris

```
iris_scaled <- scale(iris[, 1:4])
set.seed(123)
km <- kmeans(iris_scaled, centers = 3, nstart = 25)
km$size
table(cluster = km$cluster,
       species = iris$Species)

## [1] 50 53 47
##      species
## cluster setosa versicolor virginica
##      1      50           0          0
##      2       0          39         14
##      3       0          11         36
```

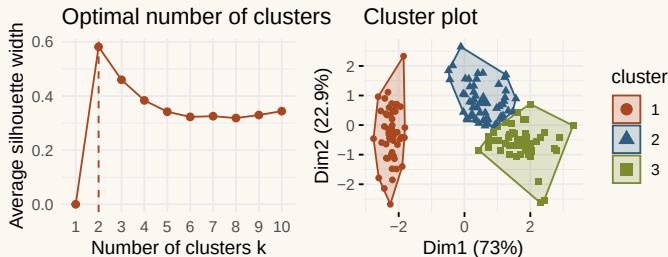
Choose $C_1, \dots, C_k$ to minimise

$$\sum_{j=1}^k \sum_{i \in C_j} \|x_i - \bar{x}_j\|^2$$

Reading it Species labels were *not* used. Setosa is recovered exactly; versicolor and virginica overlap (25 flowers in another species' cluster).

Choosing k , and seeing the clusters

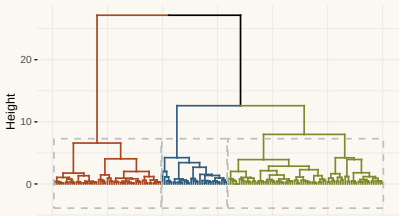
```
p_sil <- fviz_nbclust(iris_scaled, kmeans, method = "silhouette", nstart = 25,  
  linecolor = col_main) + theme_minimal()  
p_km <- fviz_cluster(km, data = iris_scaled, geom = "point", palette = unname(col_species),  
  ggtheme = theme_minimal())  
grid.arrange(p_sil, p_km, ncol = 2, widths = c(1, 1.25))
```



Reading it The silhouette peaks at $k = 2$ (setosa against the rest); knowing there are three species justifies $k = 3$.

Hierarchical clustering

```
hc <- hclust(dist(iris_scaled), method = "ward.D2")
fviz_dend(hc, k = 3, show_labels = FALSE,
          k_colors = unname(col_species),
          rect = TRUE, main = NULL,
          ggtheme = theme_minimal(base_size = 11))
```



```
table(cluster = cutree(hc, k = 3),
      species = iris$Species)
```

##	species			
##	cluster	setosa	versicolor	virginica
##	1	49	0	0
##	2	1	27	2
##	3	0	23	48

Reading it Cut at three groups, the Ward tree again isolates setosa and mixes the others.

Multiple regression for fuel economy

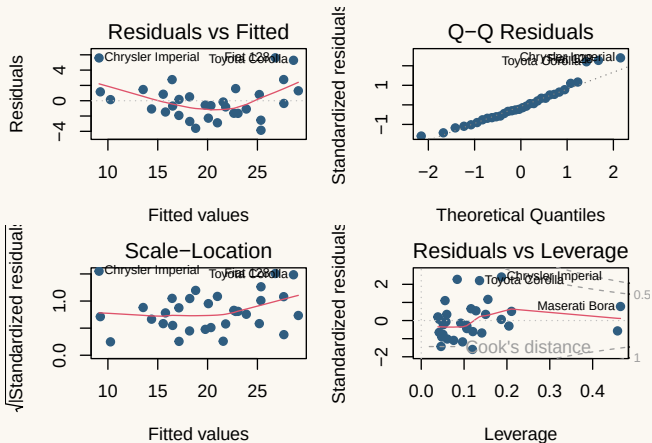
```
fit <- lm(mpg ~ wt + hp + qsec, data = mtcars)
round(coef(summary(fit)), 4)
c(R2 = summary(fit)$r.squared,
  adj_R2 = summary(fit)$adj.r.squared)
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	27.6105	8.4199	3.279	0.0028
## wt	-4.3588	0.7527	-5.791	0.0000
## hp	-0.0178	0.0150	-1.190	0.2442
## qsec	0.5108	0.4392	1.163	0.2546
## R2 adj_R2				
## 0.8348 0.8171				

$$\text{mpg}_i = \beta_0 + \beta_1 \text{wt}_i + \beta_2 \text{hp}_i + \beta_3 \text{qsec}_i + \varepsilon_i$$

Reading it $R^2 = 0.83$. Only weight is clearly significant: each extra 1,000 lb lowers mpg by 4.36, hp and qsec held fixed. hp ($p = 0.24$) and qsec ($p = 0.25$) add little once weight is in.

Regression diagnostics



```
par(mfrow = c(2, 2),
    mar = c(4, 4, 2, 1))
plot(fit, col = col_alt, pch = 19)
```

```
bptest(fit)$p.value # variance
shapiro.test(resid(fit))$p.value
round(vif(fit), 2) # collinearity

##      BP
## 0.558
## [1] 0.05539
##   wt   hp  qsec
## 2.53 4.92 2.87
```

Reading the diagnostics

- **Linearity:** residuals versus fitted bend slightly (U-shape); try $\log(\text{mpg})$ or a curved term
- **Constant variance:** Breusch–Pagan $p = 0.56$, no evidence against it
- **Normality:** Shapiro–Wilk $p = 0.055$, borderline; a few large positive residuals (Chrysler Imperial, Fiat 128, Toyota Corolla)
- **Collinearity:** VIF for hp is 4.9; hp and qsec correlate ($r = -0.71$)

Correction to the original notes

They said all three predictors were significant and the assumptions satisfied. The output shows otherwise: only weight is significant and linearity is questionable.

A simpler model, $\text{mpg} \sim \text{wt} + \text{hp}$, has adjusted $R^2 = 0.815$ (versus 0.817) with both terms significant.

Three variables at once: class, sex and survival

```
tit <- as.data.frame(Titanic)                                # one row per cell, count in Freq
tit %>% group_by(Class, Sex) %>%                             # survival rate (%) by class and sex
  summarise(rate = round(100 * sum(Freq[Survived == "Yes"]) / sum(Freq), 1)) %>%
  pivot_wider(names_from = Sex, values_from = rate) %>% as.data.frame()

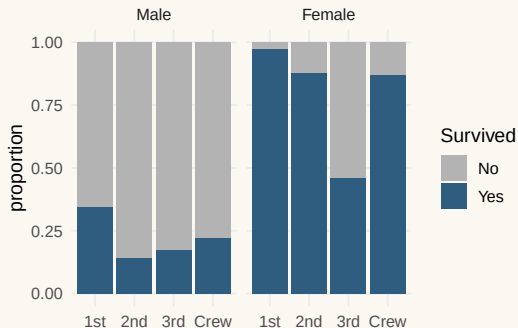
##   Class Male Female
## 1   1st 34.4   97.2
## 2   2nd 14.0   87.7
## 3   3rd 17.3   45.9
## 4  Crew 22.3   87.0
```

Reading it About 97% of first-class women survived but only 46% of third-class women, and men did worst in second class.

Class, sex and survival: the picture

```
ggplot(tit, aes(Class, weight = Freq,  
                fill = Survived)) +  
  geom_bar(position = "fill") +  
  facet_wrap(~ Sex) +  
  scale_fill_manual(values = c("grey70",  
                               col_alt)) +  
  labs(x = NULL, y = "proportion")
```

Weight by Freq Each row of `tit` is a table cell, not a person. Counting rows (as the original slide did) gives equal, meaningless bars.



Facets add a third variable to a two-variable plot. Here: 2,201 people, crew included.

PART 5

Checklist and take-aways

An EDA checklist

1. **Structure:** `dim()`, `str()`; categories as factors
2. **Quality:** missing values, blanks, duplicates
3. **One variable:** summary, plot, outliers
4. **Pairs:** plot and summary that fit the types
5. **Third variables:** colour or facet by group
6. **Many variables:** pairs plot, heatmap, PCA, clusters
7. **Model, then diagnose;** record every decision

What we found today A borderline outlier that is a real car; ages missing more often in third class; a transmission “effect” that is mostly weight; a correlation reversed by grouping; two dimensions behind eleven car variables; a regression with one useful predictor.

Check your understanding

1. Who developed and promoted EDA?
(a) George Box (b) Ronald Fisher (c) John Tukey (d) Karl Pearson
2. What does `mtcars` mainly contain?
(a) plant species (b) fuel use and car performance (c) Titanic passengers
3. Which is a *graphical* EDA technique?
(a) *t*-test (b) chi-squared test (c) histogram (d) ANOVA
4. Which is *not* a dimension of the Titanic table?
(a) Class (b) Sex (c) Survived (d) Fare
5. What does PCA mainly help with?
(a) predicting categories (b) detecting outliers (c) reducing dimension
6. Within each species, sepal length and width are correlated...
(a) negatively (b) not at all (c) positively

Answers: 1 (c), 2 (b), 3 (c), 4 (d), 5 (c), 6 (c).

Key take-aways

1. **Look first, model later:** every analysis today began with a plot, and the plots raised questions no test would have asked.
2. **Know your variable types:** they decide the plot, the summary and the test.
3. **Missing is information:** ask why values are missing before filling them.
4. **Beware of third variables:** weight explained the transmission gap; species reversed a correlation.
5. **Reduce, then interpret:** PCA and clustering summarise many variables but need scaling and judgement.
6. **Read the output, not the notes:** check every claim against the numbers and the diagnostics.

References

- Fox, J., and Weisberg, S. (2019). *An R Companion to Applied Regression*, 3rd Ed., Sage, Thousand Oaks.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*, 2nd Ed., Springer, New York.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org>
- Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, 2nd Ed., Springer, New York.
- Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G. (2023). *R for Data Science*, 2nd Ed., O'Reilly, Sebastopol. (Free online: r4ds.hadley.nz)

Thank you

Questions and discussion