

IntroSlides

Basic steps in workflow

1. *Define environment*
2. *Import*
3. *Transform*
4. *Explore (general/outlier/distribution) (go back to 3?)*
5. *Classify scale level / distribution (based on 3/4)*
6. *Describe*
7. *Test / Model (may include step 6)*
8. *Report*

Define environment

- *Activate packages to use: library() / pacman::p_load()*
- *ggplot theme: theme_set() / theme_update()*
- *flextable settings: set_flextable_defaults()*
- *knitr::opts_chunk\$set()*

```
1  if(!require(pacman)){
2    install.packages("pacman")
3  }
4  pacman::p_load(wrappedtools,
5                 ggbeeswarm, ggsignif, ggribes,
6                 car, flextable)
7
8  knitr::opts_chunk$set(message = FALSE,
9                        warning = FALSE,
10                       comment = NA)
11
12  set_flextable_defaults(
13    font.size = 9,
14    theme_fun = theme_zebra,
15    padding.bottom = 1,
16    padding.top = 3,
17    padding.left = 2,
18    padding.right = 4)
```

Import / Transform

Import

- *read_xlsx() / read_csv() / read_csv2()*
- *options related to separators, number formats, ranges etc.*

Transform

- *rename() / rename_with()*
- *mutate() / mutate(across())*
- *e.g. for log-transformation, creation of factors, text recodings*

Transformations: colnames

```
1 data(faketrial) # from wrappedtools
2 colnames(faketrial)[1:10]
```

```
[1] "Sex" "Agegroup"
[3] "Treatment" "HR"
[5] "sysRR" "diaRR"
[7] "Responder" "Med Consectetur FakePharm"
[9] "Med Sollicitudin FakePharm" "Med Suspendisse FakePharm"
```

```
1 faketrial <-
2   rename(.data = faketrial,
3         Heartrate = HR) #newname = oldname
4 faketrial <-
5   rename_with(faketrial,
6               .fn = ~str_replace(string = ., #. is placeholder
7                                pattern = "Fa.+$",
8                                replacement = "generic"))
9 colnames(faketrial)[1:10]
```

```
[1] "Sex" "Agegroup"
[3] "Treatment" "Heartrate"
[5] "sysRR" "diaRR"
[7] "Responder" "Med Consectetur generic"
[9] "Med Sollicitudin generic" "Med Suspendisse generic"
```

Transformations: content

```
1 ksnormal(faketrials$`Biomarker 1 [units]`)
```

```
[1] 0.01800259
```

```
1 faketrials <-  
2   mutate(faketrials,  
3           `Biomarker 1 ln` = log(`Biomarker 1 [units]`))  
4 ksnormal(faketrials$`Biomarker 1 ln`)
```

```
[1] 0.2090613
```

```
1 # faketrials |> select(contains('Biomarker 1 ')) |> str()  
2 faketrials <-  
3   mutate(faketrials,  
4           across(matches('Biom.+\\[\\]'),  
5                   .fns = ~.x*1000,  
6                   .names = "{.col}x1000"),  
7           across(starts_with('Med'),  
8                   .fns = factor))  
9 faketrials |>  
10  select(contains('Biomarker 1 ')) |>  
11  str()
```

```
tibble [300 × 3] (S3: tbl_df/tbl/data.frame)
```

```
$ Biomarker 1 [units]      : num [1:300] 155 151 140 130 152 ...
```

```
$ Biomarker 1 ln          : num [1:300] 5.05 5.02 4.94 4.86 5.02 ...
```

```
$ Biomarker 1 [units]x1000: num [1:300] 155306 151185 140376 129510 152014 ...
```

Explore / group variables

Explore (general/outlier/distribution)

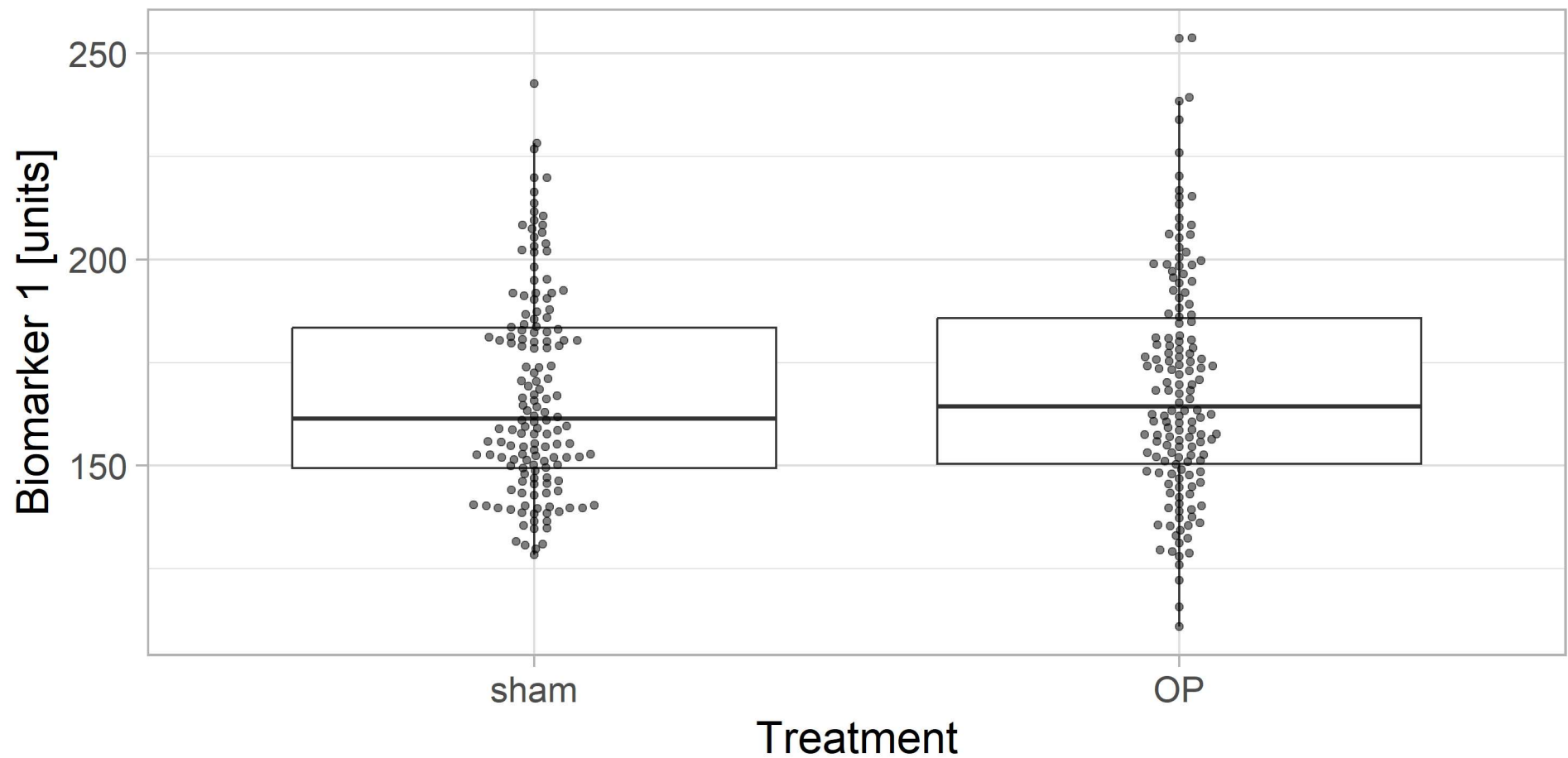
- *ggplot()+geom_boxplot() / geom_beeswarm() / geom_density()*
- *ks.test() / ksnormal() / shapiro.test()*

Classify scale level / distribution

- *gaussvars / ordvars / factvars, possibly more...*
- *Store variables accordingly, e.g. FindVars()*

Explore: Outlier

```
1 ggplot(faketrial,  
2       aes(x = Treatment,  
3           y = `Biomarker 1 [units]`))+  
4   geom_boxplot(outlier.alpha = 0) + #hide outliers, beeswarm will plot them  
5   geom_beeswarm(alpha=.5)
```



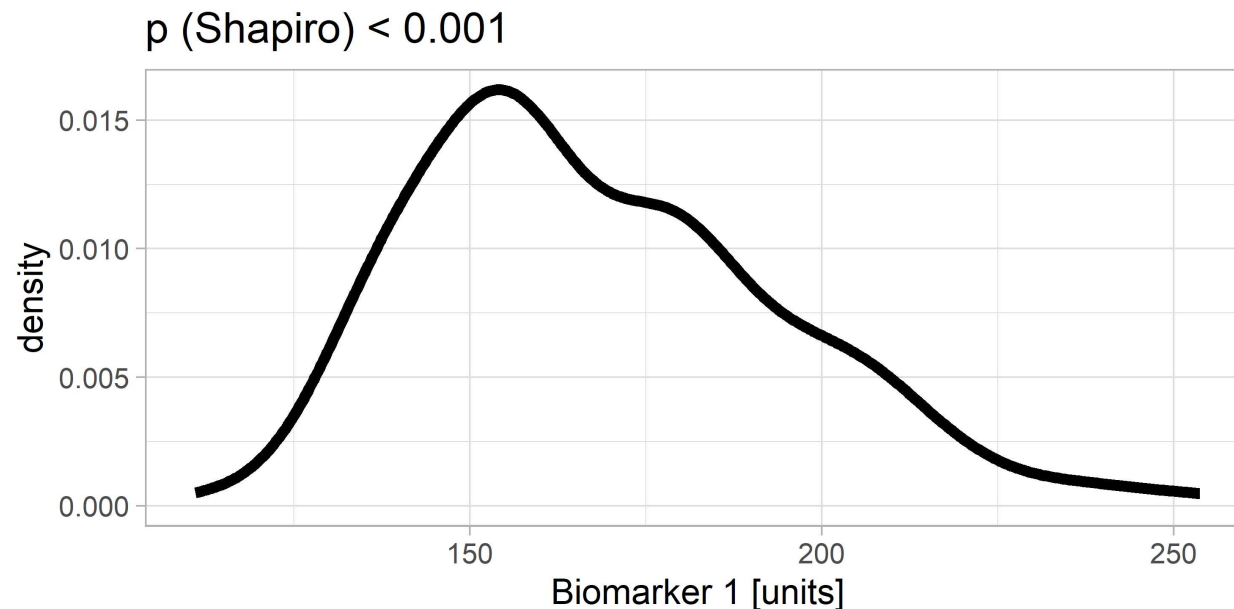
Explore: Normal distribution 1

```
1 p_normal <-  
2   shapiro.test(faketrial$`Biomarker 1 [units]`)  
3 p_normal
```

Shapiro-Wilk normality test

```
data:  faketrial$`Biomarker 1 [units]`  
W = 0.96812, p-value = 3.448e-06
```

```
1 ggplot(faketrial, aes(x = `Biomarker 1 [units]`)) +  
2   geom_density(linewidth=3) +  
3   ggtitle(paste('p (Shapiro)',  
4                 formatP(p_normal$p.value, pretext = T)))
```



Explore: Normal distribution 2

```
1 faketrial |>
2   summarize(across(.cols = starts_with('Biom'),
3                     .fns = ksnormal)) |>
4   pivot_longer(cols = everything(),
5                 names_to = 'Variable', values_to = 'pKS')
```

A tibble: 21 × 2

	Variable <chr>	pKS <dbl>
1	Biomarker 1 [units]	0.0180
2	Biomarker 2 [units]	0.479
3	Biomarker 3 [units]	0.170
4	Biomarker 4 [units]	0.976
5	Biomarker 5 [units]	0.00741
6	Biomarker 6 [units]	0.675
7	Biomarker 7 [units]	0.0873
8	Biomarker 8 [units]	0.944
9	Biomarker 9 [units]	0.150
10	Biomarker 10 [units]	0.163
# ... with 11 more rows		

Explore: Group variables

```
1 gaussvars <- FindVars(varnames = c('He','RR'),  
2                         allnames = cn(faketrial))  
3 gaussvars
```

\$index

```
[1] 4 5 6
```

\$names

```
[1] "Heartrate" "sysRR"      "diaRR"
```

\$bticked

```
[1] "`Heartrate`" "`sysRR`"      "`diaRR`"
```

\$count

```
[1] 3
```

```
1 ordvars <- FindVars(c('B'),  
2                     allnames = cn(faketrial),  
3                     exclude = c('x1','ln$'))  
4 ordvars$names |> head(n = 6)
```

```
[1] "Biomarker 1 [units]" "Biomarker 2 [units]" "Biomarker 3 [units]"
```

```
[4] "Biomarker 4 [units]" "Biomarker 5 [units]" "Biomarker 6 [units]"
```

```
1 factvars <- FindVars(c('Sex','Res','generic'),  
2                     allnames = cn(faketrial))  
3 factvars$bticked |> head(n=4)
```

```
[1] "`Sex`"          "`Responder`"
```

```
[3] "`Med Consectetur generic`" "`Med Sollicitudin generic`"
```

Model

Describe

- *mean() / sd() / meansd()*
- *median() / quantile() / median_quart()*
- *table() / prop.table() / cat_desc_stats()*

Test

- *t.test() / lm()+[Aa]nova() / compare2numvars()*
- *wilcox.test()*
- *fisher.test() / glm(family=binomial)*

Model: Describe

```
1 desc_gauss <- faketrial |>
2   summarize(across(.cols = gaussvars$names,
3                     .fns = meansd))
4 desc_gauss
```

```
# A tibble: 1 × 3
  Heartrate sysRR    diaRR
  <chr>      <chr>    <chr>
1 200 ± 22  113 ± 13 83 ± 13
```

```
1 desc_ord <- faketrial |>
2   summarize(across(ordvars$names,
3                     .fns=~median_quart(.x))) |>
4   pivot_longer(everything())
5 desc_ord
```

```
# A tibble: 10 × 2
  name                value
  <chr>              <chr>
1 Biomarker 1 [units] 163 (149/184)
2 Biomarker 2 [units] 149 (140/160)
3 Biomarker 3 [units] 164 (146/188)
4 Biomarker 4 [units] 148 (139/158)
5 Biomarker 5 [units] 163 (149/189)
6 Biomarker 6 [units] 149 (139/159)
7 Biomarker 7 [units] 167 (147/189)
8 Biomarker 8 [units] 149 (139/160)
9 Biomarker 9 [units] 165 (145/191)
10 Biomarker 10 [units] 148 (140/157)
```

Model: Test 1 / single variables

```
1 #t-Test with test for equal variances
2 t.test(formula=sysRR~Treatment, data=faketrial,
3         var.equal=var.test(formula=sysRR~Treatment,
4                             data=faketrial)$p.value>.05)
```

Welch Two Sample t-test

data: sysRR by Treatment

t = -9.4166, df = 273.41, p-value < 2.2e-16

alternative hypothesis: true difference in means between group sham and group OP is not equal to 0

95 percent confidence interval:

-15.063528 -9.854138

sample estimates:

mean in group sham	mean in group OP
107.0578	119.5166

```
1 #Wilcoxon-Test
2 wilcox.test(`Biomarker 1 [units]`~Treatment,
3             data = faketrial)
```

Wilcoxon rank sum test with continuity correction

data: Biomarker 1 [units] by Treatment

W = 10905, p-value = 0.6465

alternative hypothesis: true location shift is not equal to 0

Model: Test 2 / multiple variables

```
1 test_gauss <- compare2numvars(data = faketrials,
2                               dep_vars = gaussvars$names,
3                               indep_var = 'Treatment',
4                               gaussian = TRUE,
5                               round_p = 5)
6 test_gauss
```

```
# A tibble: 3 × 5
  Variable desc_all `Treatment sham` `Treatment OP` p
  <fct>      <chr>      <chr>      <chr>      <chr>
1 Heartrate 200 ± 22 201 ± 24      200 ± 20      0.65363
2 sysRR      113 ± 13 107 ± 13      120 ± 10      0.00001
3 diaRR       83 ± 13 76 ± 13       90 ± 9        0.00001
```

```
1 test_ord <- compare2numvars(data = faketrials,
2                              dep_vars = ordvars$names,
3                              indep_var = 'Treatment',
4                              gaussian = FALSE)
5 test_ord |> slice_head(n = 5)
```

```
# A tibble: 5 × 5
  Variable          desc_all      `Treatment sham` `Treatment OP` p
  <fct>            <chr>      <chr>      <chr>      <chr>
1 Biomarker 1 [units] 163 (149/184) 161 (149/184) 164 (150/186) 0.647
2 Biomarker 2 [units] 149 (140/160) 147 (139/159) 150 (141/160) 0.205
3 Biomarker 3 [units] 164 (146/188) 162 (147/186) 166 (146/189) 0.627
4 Biomarker 4 [units] 148 (139/158) 148 (137/158) 148 (141/158) 0.493
5 Biomarker 5 [units] 163 (149/189) 161 (147/187) 168 (150/194) 0.260
```

Model: linear models 1 / univariable

```
1 lm1<- lm(sysRR~Agegroup, data=faketrial)
2 lm1
```

Call:

```
lm(formula = sysRR ~ Agegroup, data = faketrial)
```

Coefficients:

(Intercept)	Agegroupmiddle	Agegroupold
110.669	-1.202	9.056

```
1 anova(lm1)
```

Analysis of Variance Table

Response: sysRR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Agegroup	2	6289	3144.70	21	2.95e-09 ***
Residuals	297	44476	149.75		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model: linear models 2 / multivariable

```
1 lm2<- lm(sysRR~ (Sex+Agegroup)*Treatment,  
2         data=faketrial)  
3 lm2
```

Call:

```
lm(formula = sysRR ~ (Sex + Agegroup) * Treatment, data = faketrial)
```

Coefficients:

(Intercept)	Sexmale
99.9852	2.8812
Agegroupmiddle	Agegroupold
-0.8378	18.2524
TreatmentOP	Sexmale:TreatmentOP
21.3694	-4.7366
Agegroupmiddle:TreatmentOP	Agegroupold:TreatmentOP
-1.2218	-19.0720

```
1 Anova(lm2,type = 3)
```

Anova Table (Type III tests)

Response: sysRR

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	424455	1	4632.6739	< 2.2e-16	***
Sex	303	1	3.3052	0.07009	.
Agegroup	11605	2	63.3323	< 2.2e-16	***
Treatment	8761	1	95.6165	< 2.2e-16	***
Sex:Treatment	410	1	4.4724	0.03529	*
Agegroup:Treatment	5635	2	30.7511	7.597e-13	***
Residuals	26754	292			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

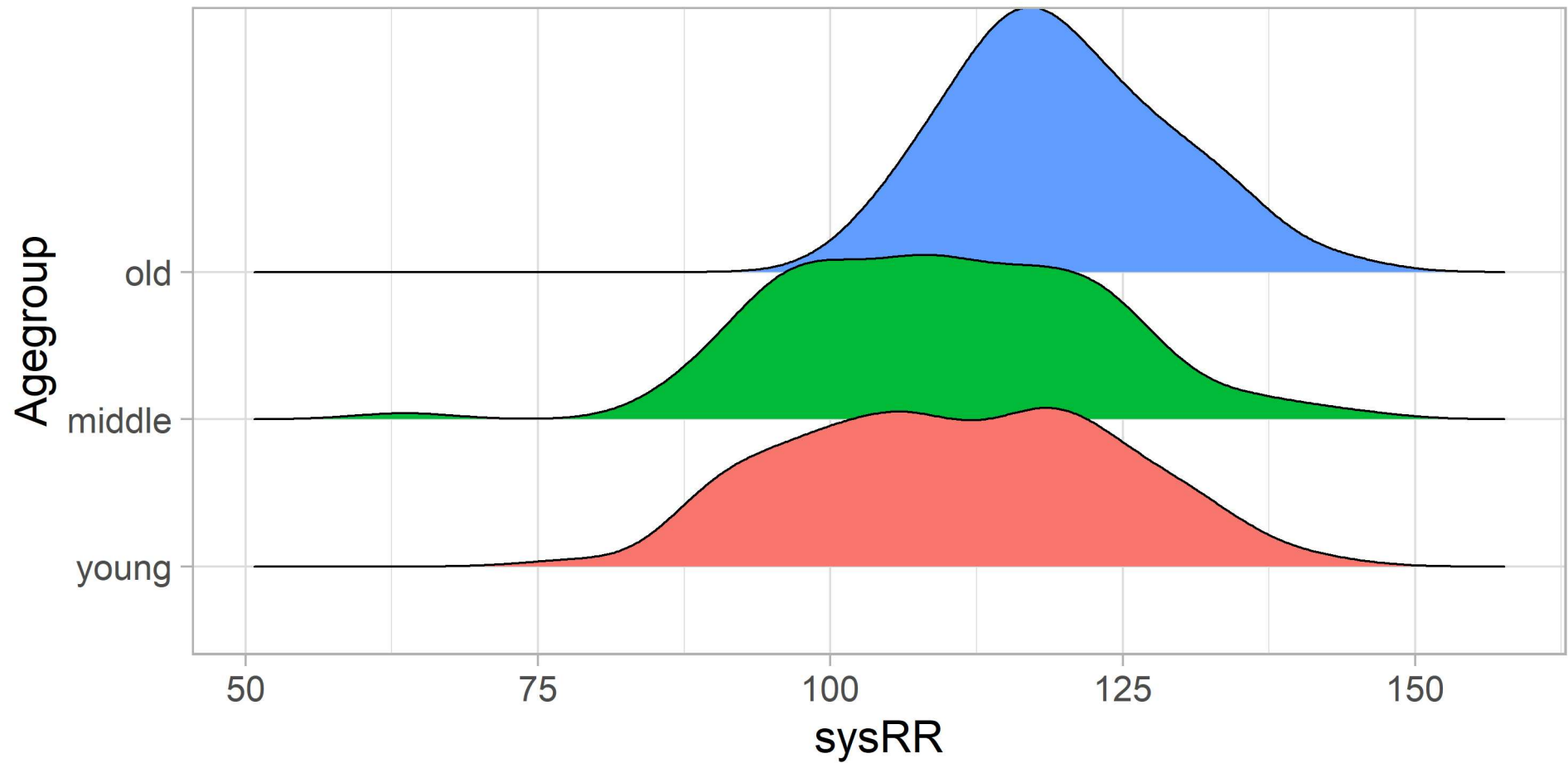
Visualize 1

```
1 ggplot(faketrial, aes(x = Treatment, y = sysRR)) +  
2   geom_violin(draw_quantiles = c(.25, .5, .75)) +  
3   geom_signif(comparisons = list(c(1, 2)),  
4               annotations =  
5                 paste('p',  
6                       formatP(test_gauss$p[2],  
7                               pretext = T))) +  
8   scale_y_continuous(expand = expansion(mult = .1))
```



Visualize 2

```
1 ggplot(faketrial, aes(x = sysRR, y = Agegroup, fill=Agegroup))+  
2   geom_density_ridges()+  
3   guides(fill='none')
```



Report

- *RMarkdown and quarto are powerful tools to create reports*
- *Package flextable provides nice features for table formatting*

```
1 test_gauss |>
2   flextable() |>
3   bg(~p<=0.05,j = 5,bg = 'yellow') |>
4   set_caption('Treatment effects, measures following a normal distribution') |>
5   add_footer_lines('Significance level is set at .0.5')
```

Treatment effects, Measures following a normal distribution				
Variable	desc_all	Treatment sham	Treatment OP	p
Heartrate	200 ± 22	201 ± 24	200 ± 20	0.65363
sysRR	113 ± 13	107 ± 13	120 ± 10	0.00001
diaRR	83 ± 13	76 ± 13	90 ± 9	0.00001

Significance level is set at .0.5