CarrotTransformation

Pankaj Bhatta

04/05/2021

Necessary items:

- **libraries:** tidyverse, ggsci, ggforce, patchwork, Hmisc
- **files:** carrot_df.csv

Within RStudio you can directly edit these blocks of code and turn in your homework that way with your code and figures printed.
I **HIGHLY** recommend getting your code to work in a normal R script then just copy and pasting the final code over to this document

First: import libraries, set your working directory, and read in bloom_df

```
library(tidyverse)
library(ggforce)
library(ggsci)
library(patchwork)
library(Hmisc)
setwd('~/Desktop/BIOL792_2') #change to match your ggplot directory
carrot_df <- read.csv('carrot_transformation_version2-1.csv')
```

carrot_df contents

- **carrot_variety**
- **carrot_parts**
- **agrobacteria_type**
- **gene**
- **conditions**
- **time_in_MS1D_dark**
- **time_in_MS1D_dl**
- **time_in_half_MS1D_dl**
- **time_in_quarter_MS1D_dl**
- **antibiotics_in_MS1D_dl**

- **antibiotics_in_half_MS1D_dl**
- **antibiotics_in_qaurter_MS1D_dl**
- **callus_status**
- **PCR_Gel_status**
- **GC_MS_polyacetylene_percentage**

Let's take a peak and look at the structure

```
carrot_df[sample(nrow(carrot_df), 5), ]
```

```
##        carrot_variety carrot_parts agrobacteria_type         gene
conditions
## 35  danvers_half_long      petiole           LBA4404     acet_552
cocultivated
## 297 danvers_half_long         stem            EHA105 peaq_p19_gfp
cocultivated
## 293 danvers_half_long         root            EHA105 peaq_p19_gfp
cocultivated
## 237 danvers_half_long         leaf            EHA105     acet_552
cocultivated
## 69  danvers_half_long         root           LBA4404 empty_vector
cocultivated
##     time_in_MS1D_dark time_in_MS1D_dl time_in_half_MS1D_dl
## 35                  3              14                    56
## 297                 3              14                    56
## 293                 3              14                    56
## 237                 3              14                    56
## 69                  3              14                    56
##     time_in_quarter_MS1D_dl antibiotics_in_MS1D_dl
antibiotics_in_half_MS1D_dl
## 35                       14              required
required
## 297                      14              required
required
## 293                      14              required
required
## 237                      14              required
required
## 69                       14              required
required
##     antibiotics_in_qaurter_MS1D_dl callus_status PCR_Gel_status
## 35                        required       present        present
## 297                       required       present        present
## 293                       required       present        present
## 237                       required       present        present
```

```
## 69                          required       present        present
##     GC_MS_polyacetylene_percentage
## 35                                92
## 297                               10
## 293                               12
## 237                               72
## 69                                12

summary(carrot_df)

##  carrot_variety     carrot_parts       agrobacteria_type      gene
##  Length:320         Length:320         Length:320         Length:320
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   conditions        time_in_MS1D_dark time_in_MS1D_dl time_in_half_MS1D_dl
##  Length:320         Min.   :3         Min.   :14      Min.   :56
##  Class :character   1st Qu.:3         1st Qu.:14      1st Qu.:56
##  Mode  :character   Median :3         Median :14      Median :56
##                     Mean   :3         Mean   :14      Mean   :56
##                     3rd Qu.:3         3rd Qu.:14      3rd Qu.:56
##                     Max.   :3         Max.   :14      Max.   :56
##  time_in_quarter_MS1D_dl antibiotics_in_MS1D_dl
antibiotics_in_half_MS1D_dl
##  Min.   :14              Length:320             Length:320
##  1st Qu.:14              Class :character       Class :character
##  Median :14              Mode  :character       Mode  :character
##  Mean   :14
##  3rd Qu.:14
##  Max.   :14
##  antibiotics_in_qaurter_MS1D_dl callus_status      PCR_Gel_status
##  Length:320                     Length:320         Length:320
##  Class :character               Class :character   Class :character
##  Mode  :character               Mode  :character   Mode  :character
##
##
##
##  GC_MS_polyacetylene_percentage
##  Min.   : 0.00
##  1st Qu.:12.00
##  Median :19.00
##  Mean   :31.68
##  3rd Qu.:69.00
##  Max.   :92.00
```

We are creating ggplot of the following data for the visualization.

The layout will be:

- x axis: carrot_variety/carrot_parts/agrobacteria_type/gene
- y axis: value of GC_MS_polyacetylene_percentage

We are creating two plots for visualization:

- bar and error bars (mean and 95% conf. int.)
- raw data + point and error bars (mean and 95% conf. int.)

```
carrot_long_df <- carrot_df %>%
  gather(key=polyacetylene, value = mean, c(GC_MS_polyacetylene_percentage))
```

Calculation and summarization of polyacetylene based on carrot_variety

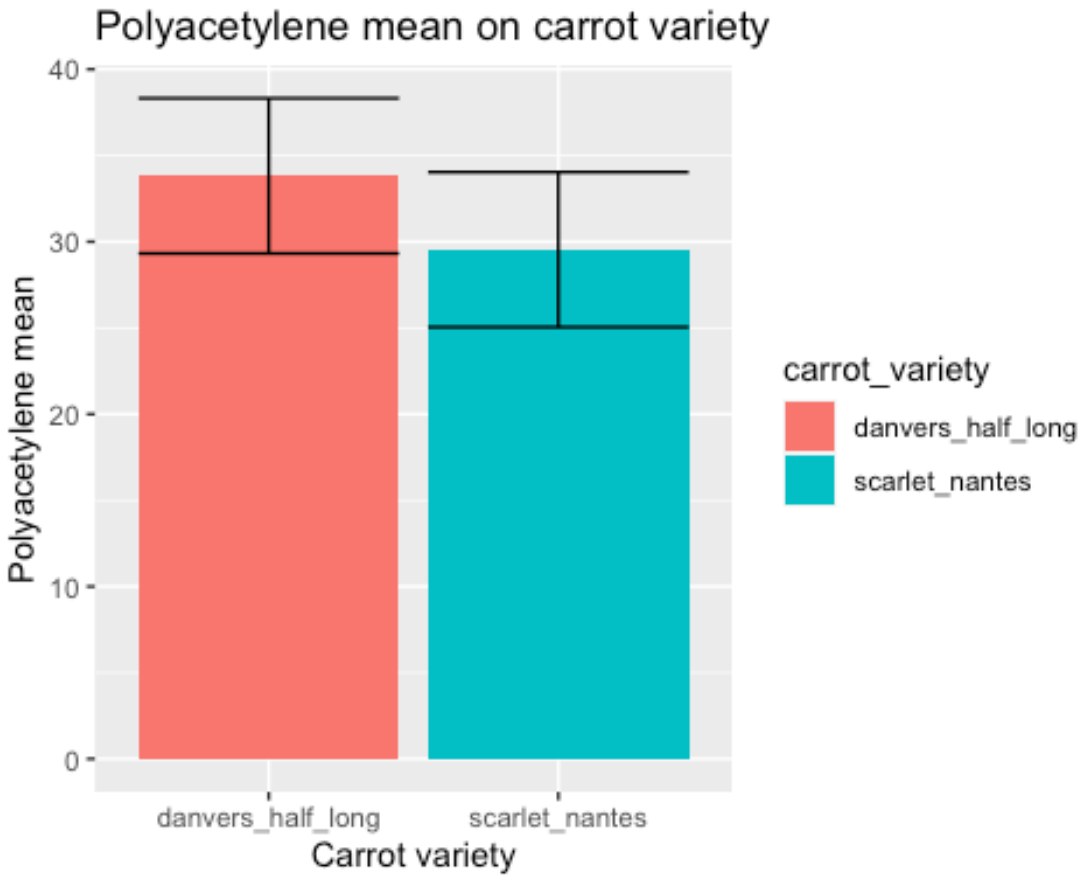
```
variety_polyacetylene_sum_df <- carrot_df %>%
  group_by(carrot_variety) %>%
  summarise(mean = mean(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            sd = sd(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            n = n()) %>%
  mutate(se = sd / sqrt(n),
         ci = 1.96*se)
variety_polyacetylene_sum_df
```

```
## # A tibble: 2 x 6
##   carrot_variety     mean    sd     n    se    ci
##   <chr>             <dbl> <dbl> <int> <dbl> <dbl>
## 1 danvers_half_long  33.8  29.1   160  2.30  4.50
## 2 scarlet_nantes     29.5  29.0   160  2.29  4.49
```

Bar and error bars (mean and 95% conf. int.)

```
carrot_variety_bar <- ggplot(data=variety_polyacetylene_sum_df,
aes(x=carrot_variety, y=mean, fill= carrot_variety))+
  geom_bar(stat = 'identity')+
  geom_errorbar(aes(ymin = mean - ci, ymax = mean + ci))+
  xlab('Carrot variety')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene mean on carrot variety')

carrot_variety_bar
```

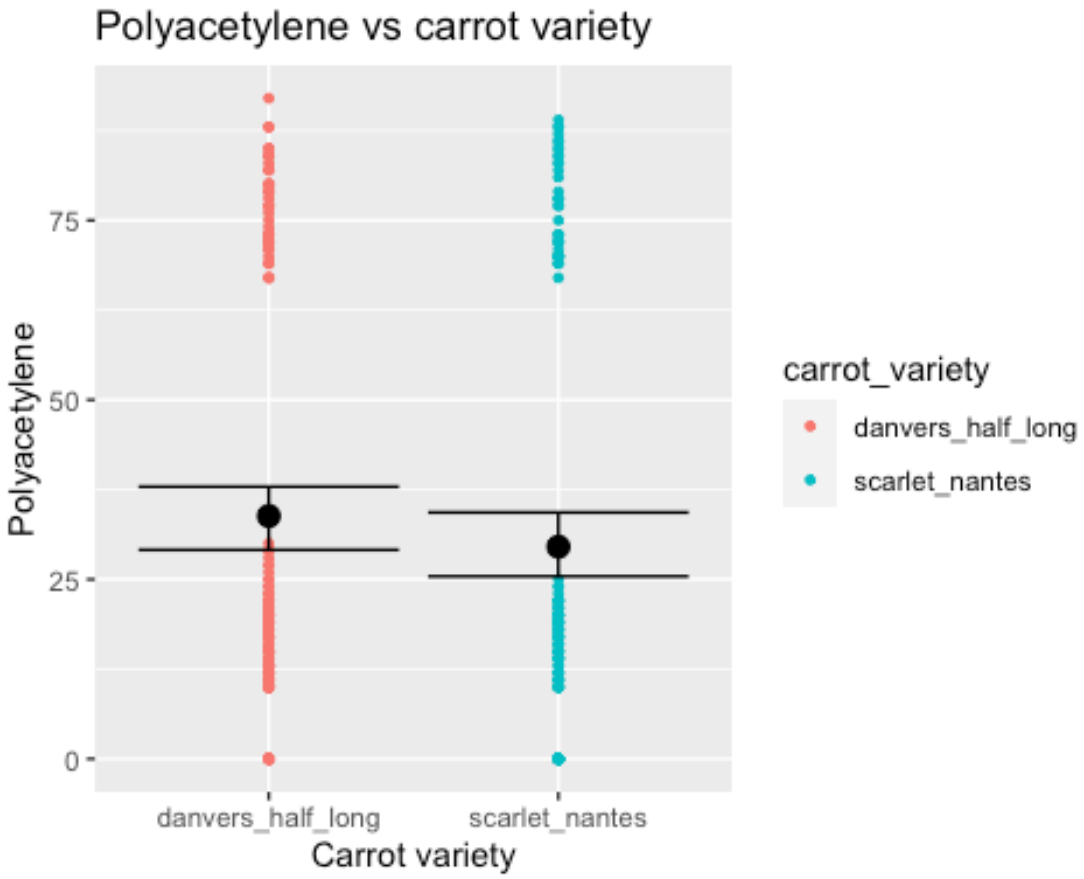


####

Raw data + point and error bars (mean and 95% conf. int.)

```
carrot_variety_plots = ggplot(data=carrot_long_df, aes(x=carrot_variety,
y=mean, colour=carrot_variety))+
  geom_point(size=1)+
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", color='black')+
  stat_summary(fun = mean, geom = "point", size=3, color='black')+
  xlab('Carrot variety')+
  ylab('Polyacetylene')+
  ggtitle('Polyacetylene vs carrot variety')

carrot_variety_plots
```



####calculation and summarization of polyacetylene based on carrot_parts

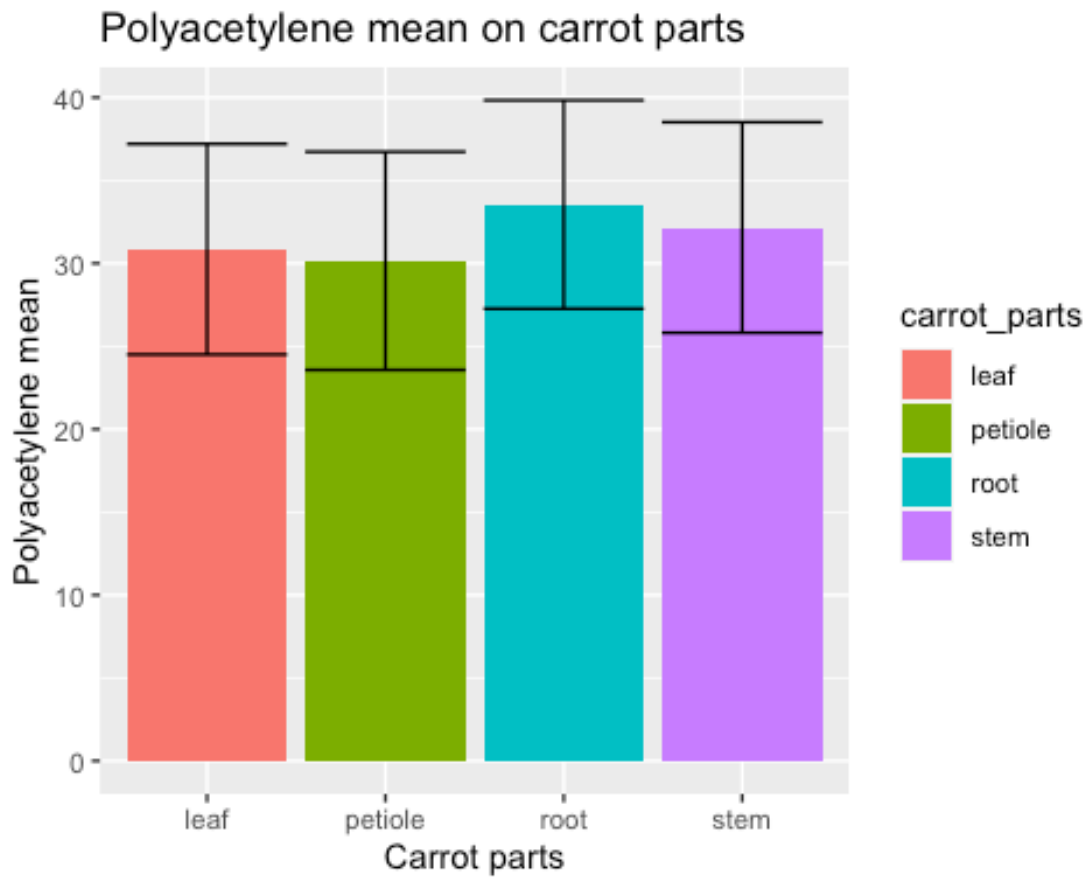
```
parts_polyacetylene_sum_df <- carrot_df %>%
  group_by(carrot_parts) %>%
  summarise(mean = mean(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            sd = sd(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            n = n()) %>%
  mutate(se = sd / sqrt(n),
         ci = 1.96*se)
parts_polyacetylene_sum_df
```

```
## # A tibble: 4 x 6
##   carrot_parts  mean    sd     n    se    ci
##   <chr>        <dbl> <dbl> <int> <dbl> <dbl>
## 1 leaf          30.8  29.0    80  3.24  6.35
## 2 petiole       30.2  30.0    80  3.36  6.58
## 3 root          33.6  28.7    80  3.21  6.29
## 4 stem          32.2  29.0    80  3.24  6.35
```

Bar and error bars (mean and 95% conf. int.)

```
carrot_parts_bar <- ggplot(data=parts_polyacetylene_sum_df,
aes(x=carrot_parts, y=mean, fill= carrot_parts))+
  geom_bar(stat = 'identity')+
  geom_errorbar(aes(ymin = mean - ci, ymax = mean + ci))+
```

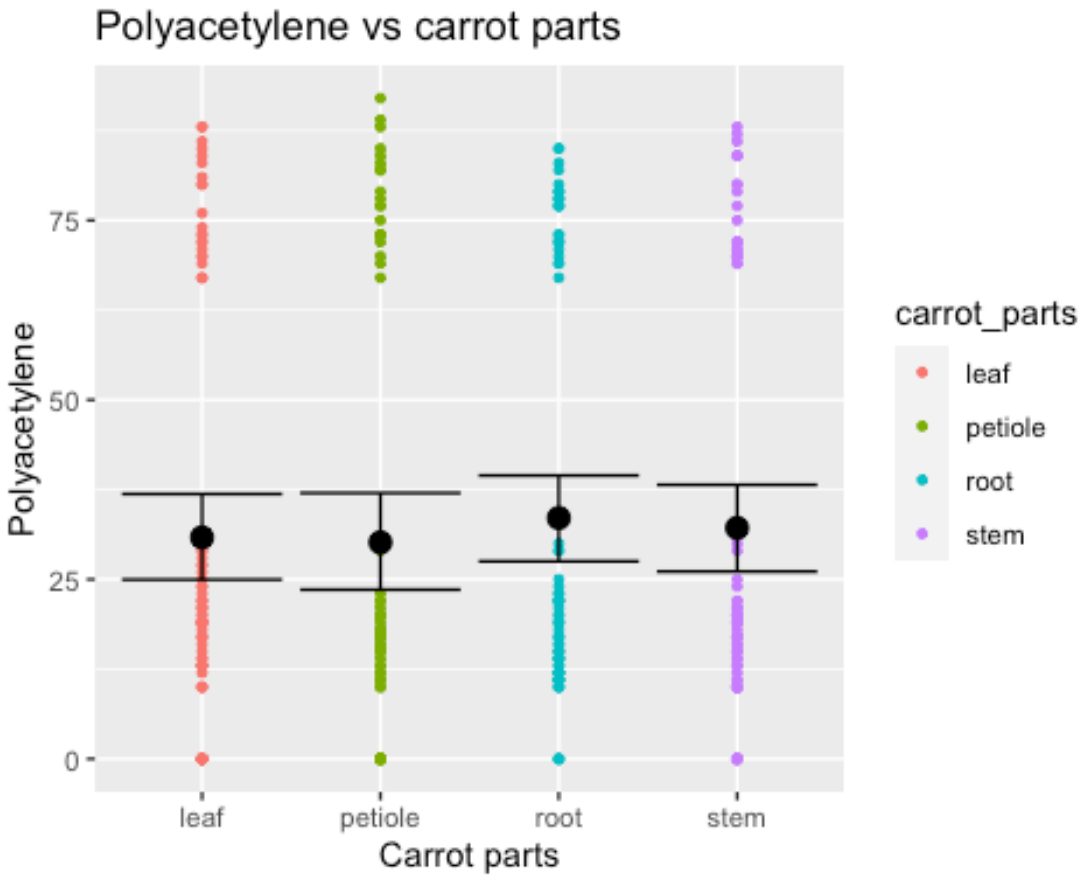
```
  xlab('Carrot parts')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene mean on carrot parts')
carrot_parts_bar
```



Raw data + point and error bars (mean and 95% conf. int.)

```
carrot_parts_plots = ggplot(data=carrot_long_df, aes(x=carrot_parts, y=mean,
colour=carrot_parts))+
  geom_point(size=1)+
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", color='black')+
  stat_summary(fun = mean, geom = "point", size=3, color='black')+
  xlab('Carrot parts')+
  ylab('Polyacetylene')+
  ggtitle('Polyacetylene vs carrot parts')

carrot_parts_plots
```



Calculation and summarization of polyacetylene based on agrobacteria_type

```
agrobacteria_polyacetylene_sum_df <- carrot_df %>%
  group_by(agrobacteria_type) %>%
  summarise(mean = mean(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            sd = sd(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            n = n()) %>%
  mutate(se = sd / sqrt(n),
         ci = 1.96*se)
agrobacteria_polyacetylene_sum_df
```

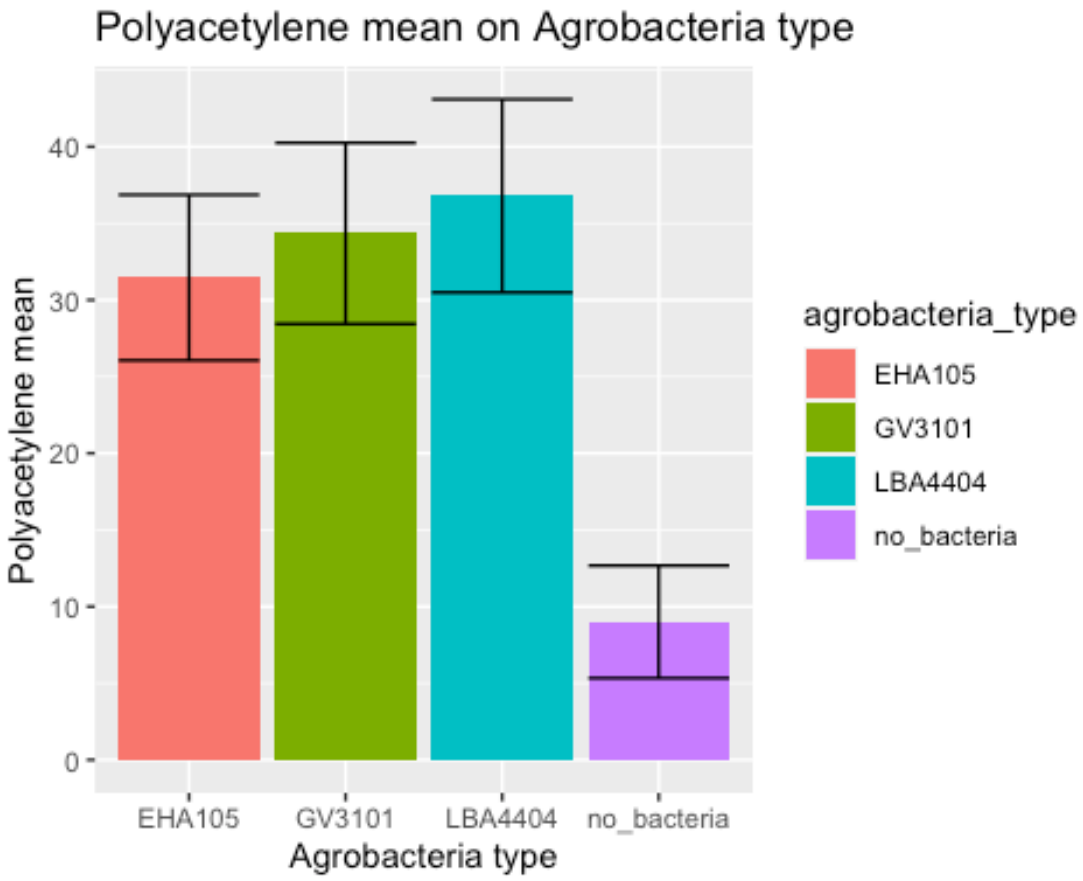
```
## # A tibble: 4 x 6
##   agrobacteria_type  mean    sd     n    se    ci
##   <chr>             <dbl> <dbl> <int> <dbl> <dbl>
## 1 EHA105             31.5  27.0    96  2.76  5.40
## 2 GV3101             34.3  29.5    96  3.01  5.90
## 3 LBA4404            36.8  31.5    96  3.22  6.30
## 4 no_bacteria         9    10.6    32  1.87  3.67
```

Bar and error bars (mean and 95% conf. int.)

```
agrobacteria_bar <- ggplot(data=agrobacteria_polyacetylene_sum_df,
aes(x=agrobacteria_type, y=mean, fill= agrobacteria_type))+
  geom_bar(stat = 'identity')+
  geom_errorbar(aes(ymin = mean - ci, ymax = mean + ci))+
```

```
  xlab('Agrobacteria type')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene mean on Agrobacteria type')

agrobacteria_bar
```

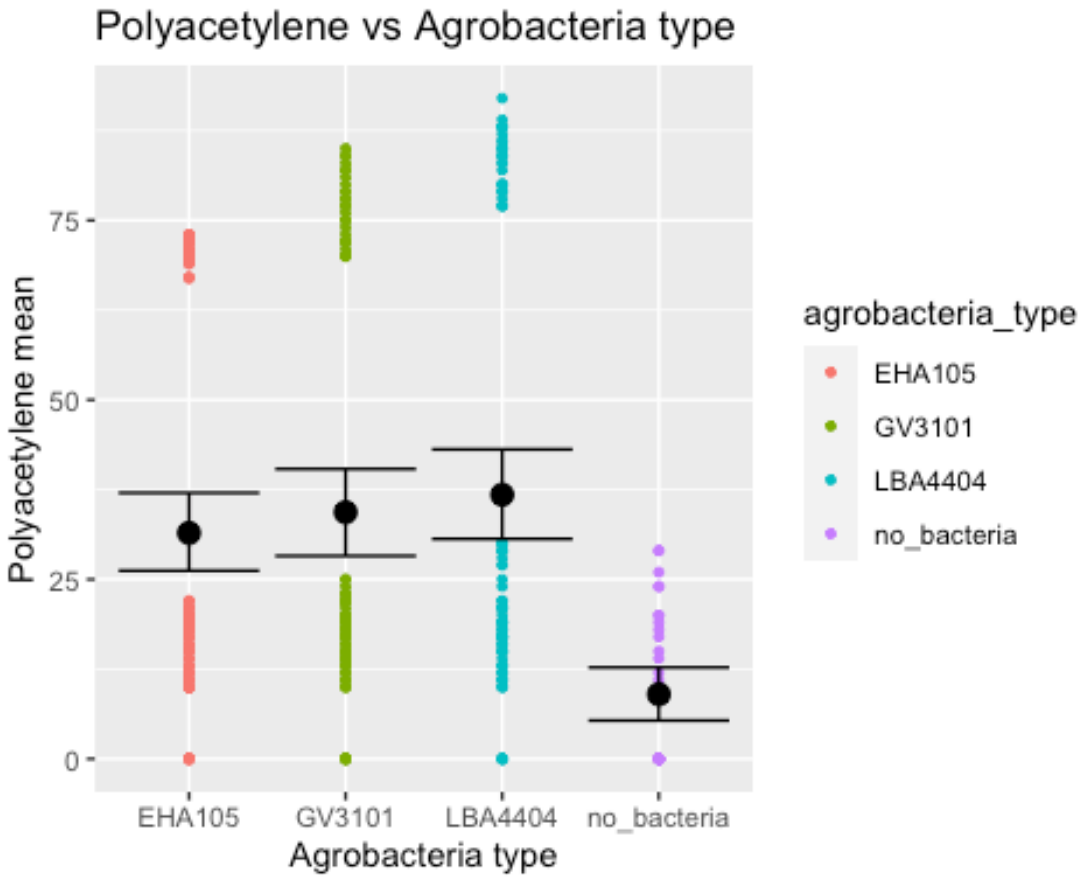


####

Raw data + point and error bars (mean and 95% conf. int.)

```
agrobacteria_plots = ggplot(data=carrot_long_df, aes(x=agrobacteria_type,
y=mean, colour=agrobacteria_type))+
  geom_point(size=1)+
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", color='black')+
  stat_summary(fun = mean, geom = "point", size=3, color='black')+
  xlab('Agrobacteria type')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene vs Agrobacteria type')

agrobacteria_plots
```



Calculation and summarization of polyacetylene based on gene

```
gene_polyacetylene_sum_df <- carrot_df %>%
  group_by(gene) %>%
  summarise(mean = mean(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            sd = sd(GC_MS_polyacetylene_percentage, na.rm= TRUE),
            n = n()) %>%
  mutate(se = sd / sqrt(n),
         ci = 1.96*se)
gene_polyacetylene_sum_df
```

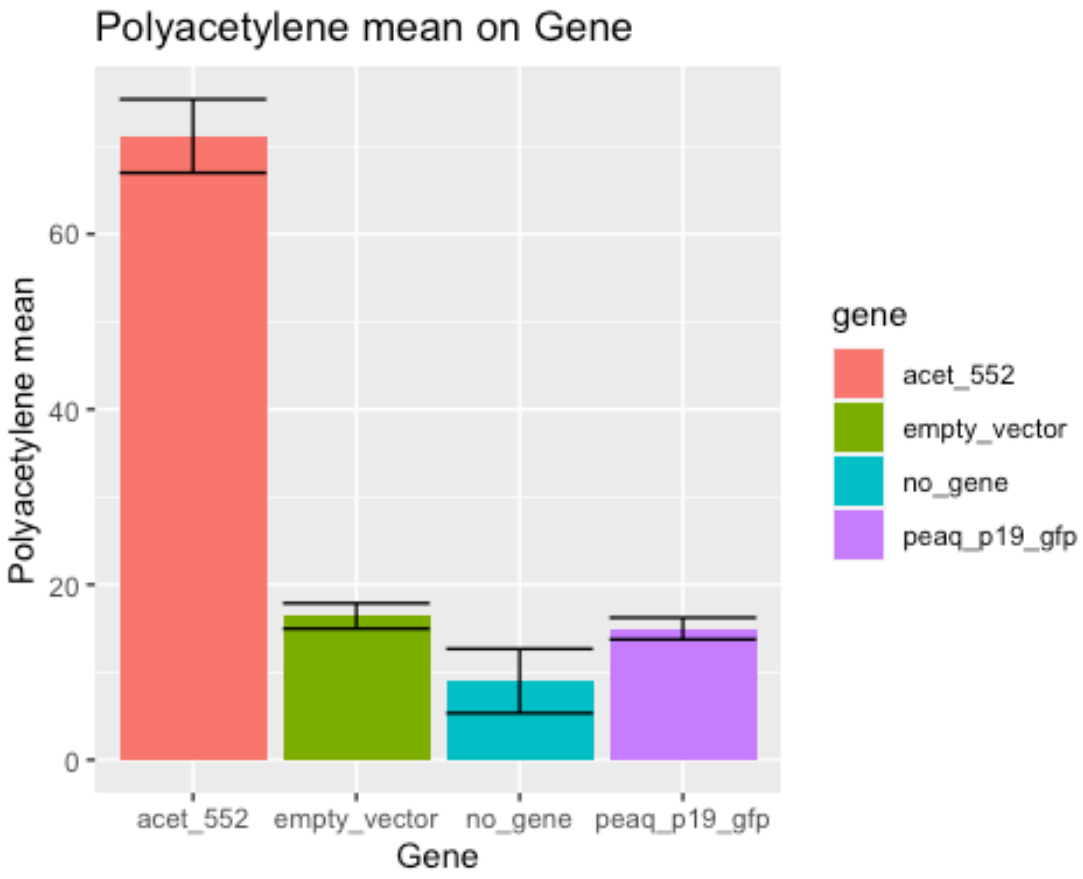
```
## # A tibble: 4 x 6
##   gene          mean    sd     n    se    ci
##   <chr>        <dbl> <dbl> <int> <dbl> <dbl>
## 1 acet_552      71.2 21.0     96 2.14   4.20
## 2 empty_vector  16.4  7.24    96 0.739  1.45
## 3 no_gene        9   10.6     32 1.87   3.67
## 4 peaq_p19_gfp  15.0  6.18    96 0.631  1.24
```

Bar and error bars (mean and 95% conf. int.)

```
gene_bar <- ggplot(data=gene_polyacetylene_sum_df, aes(x=gene, y=mean, fill=
gene))+
  geom_bar(stat = 'identity')+
  geom_errorbar(aes(ymin = mean - ci, ymax = mean + ci))+
```

```
  xlab('Gene')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene mean on Gene')

gene_bar
```

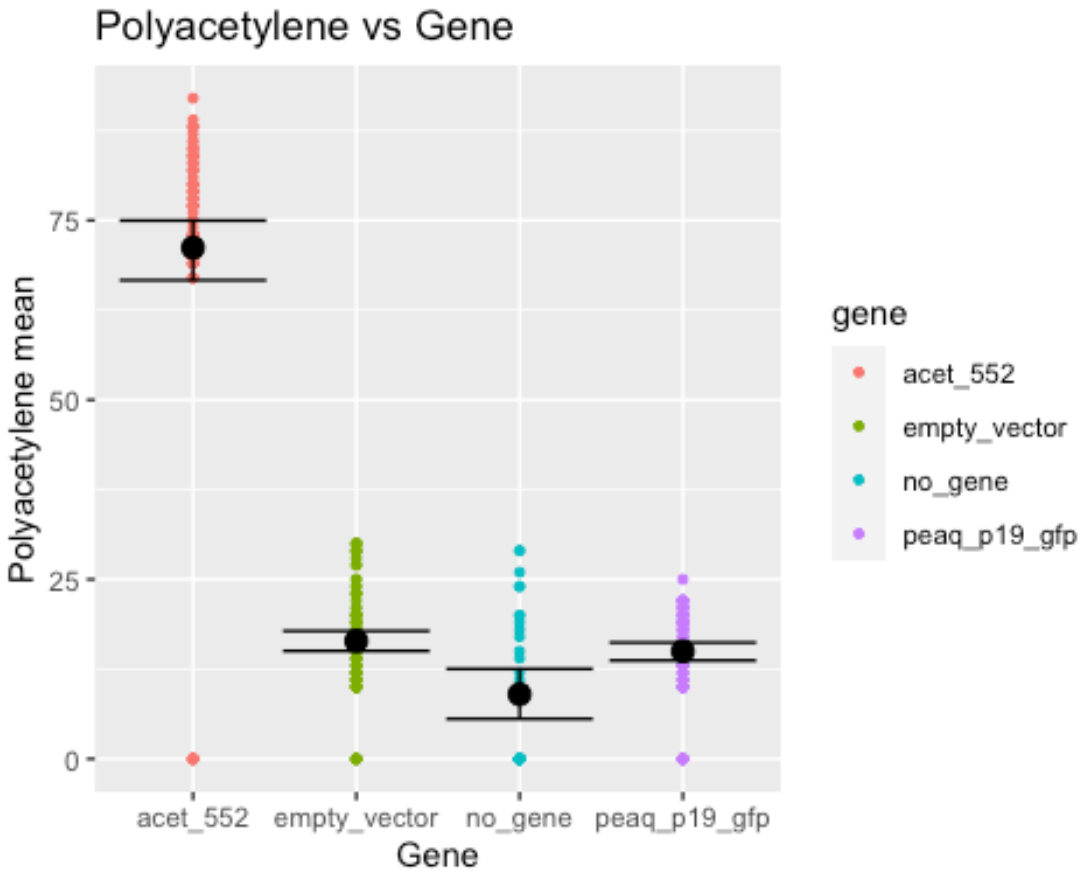


####

Raw data + point and error bars (mean and 95% conf. int.)

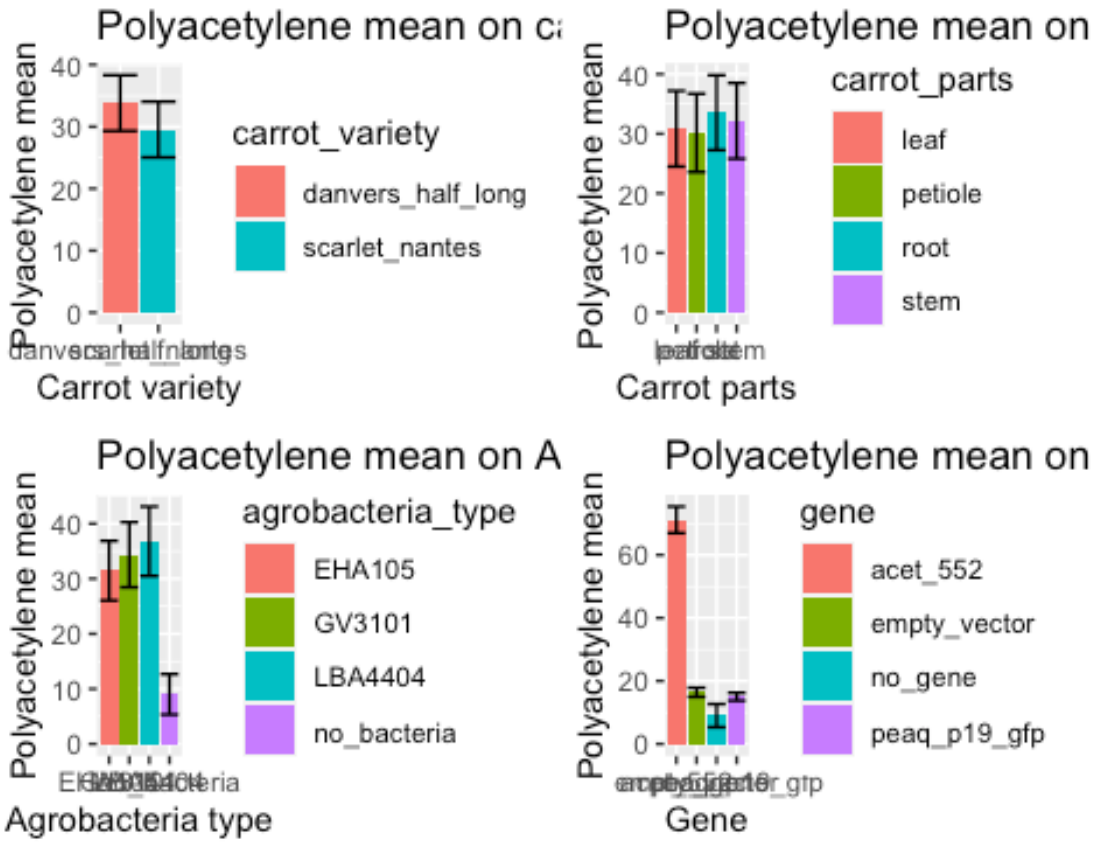
```
gene_plots = ggplot(data=carrot_long_df, aes(x=gene, y=mean, colour=gene))+
  geom_point(size=1)+
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", color='black')+
  stat_summary(fun = mean, geom = "point", size=3, color='black')+
  xlab('Gene')+
  ylab('Polyacetylene mean')+
  ggtitle('Polyacetylene vs Gene')

gene_plots
```



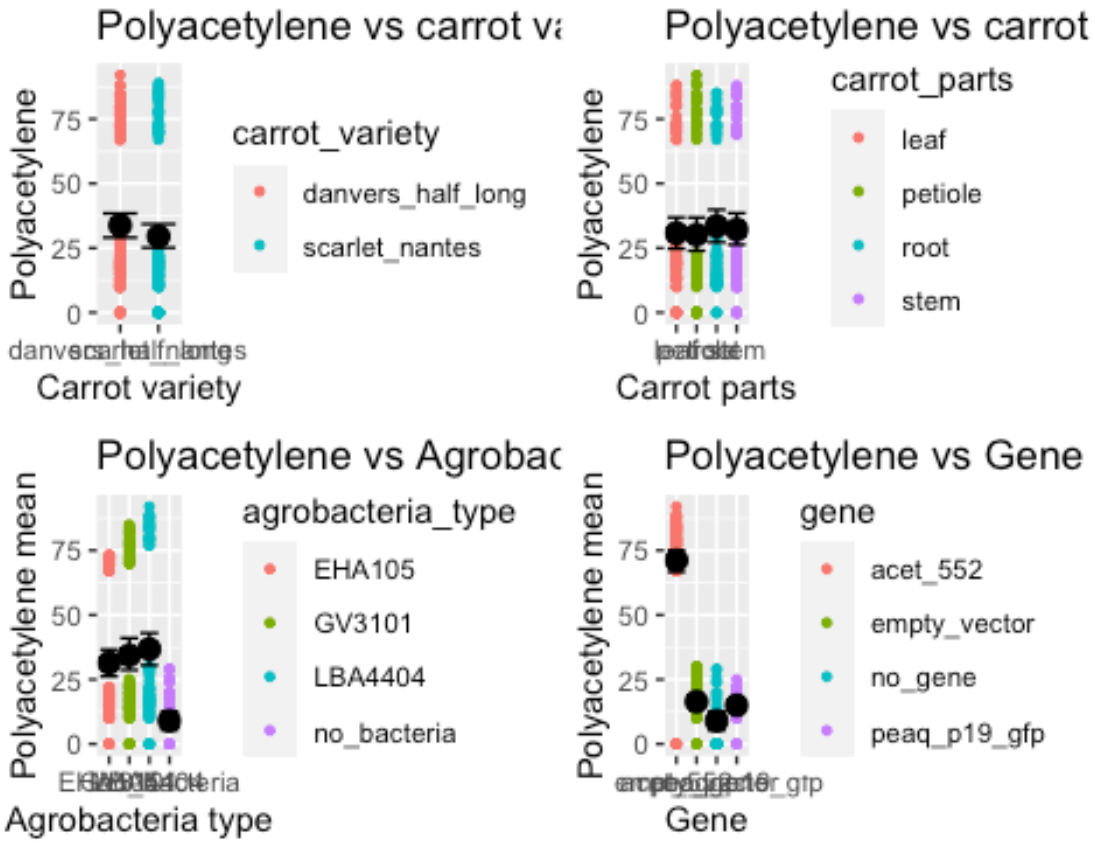
Summarizing bar plots

```
carrot_variety_bar + carrot_parts_bar + agrobacteria_bar + gene_bar
```



Summarizing point plots

```
carrot_variety_plots + carrot_parts_plots + agrobacteria_plots + gene_plots
```



Linear regression model

```
smp_siz = floor(0.75*nrow(carrot_df))
set.seed(123)
train_ind = sample(seq_len(nrow(carrot_df)),size = smp_siz)
train = carrot_df[train_ind,]
test = carrot_df[-train_ind,]
linear_model = lm(carrot_df$GC_MS_polyacetylene_percentage ~
carrot_df$carrot_variety + carrot_df$carrot_parts +
carrot_df$agrobacteria_type + carrot_df$gene, data = train)
summary(linear_model)

##
## Call:
## lm(formula = carrot_df$GC_MS_polyacetylene_percentage ~
carrot_df$carrot_variety +
##     carrot_df$carrot_parts + carrot_df$agrobacteria_type + carrot_df$gene,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.125  -3.451   2.001   6.067  22.969
##
## Coefficients: (1 not defined because of singularities)
```

```
##                                      Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                            69.760      2.208  31.595  < 2e-
16 ***
## carrot_df$carrot_varietyscarlet_nantes   -4.281      1.429  -2.997
0.00295 **
## carrot_df$carrot_partspetiole            -0.700      2.020  -0.346
0.72921
## carrot_df$carrot_partsroot                2.700      2.020   1.336
0.18238
## carrot_df$carrot_partsstem                1.312      2.020   0.650
0.51639
## carrot_df$agrobacteria_typeGV3101         2.885      1.844   1.565
0.11871
## carrot_df$agrobacteria_typeLBA4404        5.333      1.844   2.892
0.00410 **
## carrot_df$agrobacteria_typeno_bacteria  -59.448      2.817 -21.102  < 2e-
16 ***
## carrot_df$geneempty_vector              -54.771      1.844 -29.698  < 2e-
16 ***
## carrot_df$geneno_gene                        NA         NA      NA
NA
## carrot_df$genepeaq_p19_gfp              -56.198      1.844 -30.472  < 2e-
16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.78 on 310 degrees of freedom
## Multiple R-squared:  0.8121, Adjusted R-squared:  0.8067
## F-statistic: 148.9 on 9 and 310 DF,  p-value: < 2.2e-16
```

Prediction in R using linear regression

```
precentage_prediction = predict(linear_model, newdata = test)

## Warning: 'newdata' had 80 rows but variables found have 320 rows

## Warning in predict.lm(linear_model, newdata = test): prediction from a
rank-
## deficient fit may be misleading

actuals_preds <-
data.frame(cbind(actuals=carrot_df$GC_MS_polyacetylene_percentage,
predicteds=precentage_prediction))
actuals_preds

##     actuals predicteds
## 1        20    9.61250
## 2        15    9.61250
## 3         0    9.61250
## 4         0    9.61250
## 5        12   13.01250
```

```
## 6        14   13.01250
## 7         0   13.01250
## 8         0   13.01250
## 9         0   11.62500
## 10       24   11.62500
## 11        0   11.62500
## 12        0   11.62500
## 13       26   10.31250
## 14       29   10.31250
## 15        0   10.31250
## 16        0   10.31250
## 17       18    5.33125
## 18       17    5.33125
## 19        0    5.33125
## 20        0    5.33125
## 21       11    8.73125
## 22       10    8.73125
## 23        0    8.73125
## 24        0    8.73125
## 25       20    7.34375
## 26       19    7.34375
## 27        0    7.34375
## 28        0    7.34375
## 29       24    6.03125
## 30       29    6.03125
## 31        0    6.03125
## 32        0    6.03125
## 33       82   74.39375
## 34       77   74.39375
## 35       92   74.39375
## 36       79   74.39375
## 37       79   77.79375
## 38       85   77.79375
## 39       77   77.79375
## 40       80   77.79375
## 41       88   76.40625
## 42       80   76.40625
## 43       84   76.40625
## 44       79   76.40625
## 45       84   75.09375
## 46       88   75.09375
## 47       80   75.09375
## 48       80   75.09375
## 49       83   70.11250
## 50       88   70.11250
## 51       89   70.11250
## 52        0   70.11250
## 53       77   73.51250
## 54       83   73.51250
## 55       78   73.51250
```

```
## 56       85   73.51250
## 57       86   72.12500
## 58       87   72.12500
## 59        0   72.12500
## 60       84   72.12500
## 61       86   70.81250
## 62       88   70.81250
## 63        0   70.81250
## 64       85   70.81250
## 65       13   19.62292
## 66       11   19.62292
## 67       17   19.62292
## 68       18   19.62292
## 69       12   23.02292
## 70       29   23.02292
## 71       13   23.02292
## 72       17   23.02292
## 73       17   21.63542
## 74       10   21.63542
## 75       15   21.63542
## 76       21   21.63542
## 77       30   20.32292
## 78       28   20.32292
## 79       27   20.32292
## 80       27   20.32292
## 81        0   15.34167
## 82       29   15.34167
## 83       17   15.34167
## 84        0   15.34167
## 85       29   18.74167
## 86       16   18.74167
## 87       24   18.74167
## 88       30   18.74167
## 89       29   17.35417
## 90       15   17.35417
## 91       30   17.35417
## 92       15   17.35417
## 93       30   16.04167
## 94       20   16.04167
## 95        0   16.04167
## 96       19   16.04167
## 97       22   18.19583
## 98       18   18.19583
## 99       10   18.19583
## 100      21   18.19583
## 101      19   21.59583
## 102      11   21.59583
## 103      15   21.59583
## 104      12   21.59583
## 105      11   20.20833
```

```
## 106      19   20.20833
## 107      16   20.20833
## 108      13   20.20833
## 109      21   18.89583
## 110      13   18.89583
## 111      25   18.89583
## 112      18   18.89583
## 113      17   13.91458
## 114      15   13.91458
## 115      19   13.91458
## 116      22   13.91458
## 117      11   17.31458
## 118      15   17.31458
## 119      22   17.31458
## 120      12   17.31458
## 121      10   15.92708
## 122      17   15.92708
## 123      16   15.92708
## 124      14   15.92708
## 125       0   14.61458
## 126      17   14.61458
## 127      19   14.61458
## 128      21   14.61458
## 129      78   71.94583
## 130      82   71.94583
## 131      75   71.94583
## 132      85   71.94583
## 133      79   75.34583
## 134      82   75.34583
## 135      72   75.34583
## 136      77   75.34583
## 137      80   73.95833
## 138      84   73.95833
## 139      77   73.95833
## 140      71   73.95833
## 141      76   72.64583
## 142      74   72.64583
## 143      83   72.64583
## 144      72   72.64583
## 145      82   67.66458
## 146      77   67.66458
## 147      84   67.66458
## 148       0   67.66458
## 149      78   71.06458
## 150      78   71.06458
## 151      70   71.06458
## 152      79   71.06458
## 153      72   69.67708
## 154      70   69.67708
## 155      70   69.67708
```

```
## 156      75   69.67708
## 157      81   68.36458
## 158      73   68.36458
## 159       0   68.36458
## 160      70   68.36458
## 161      20   17.17500
## 162      23   17.17500
## 163      17   17.17500
## 164      11   17.17500
## 165      20   20.57500
## 166      17   20.57500
## 167      23   20.57500
## 168      25   20.57500
## 169      15   19.18750
## 170      14   19.18750
## 171      10   19.18750
## 172      13   19.18750
## 173      13   17.87500
## 174      17   17.87500
## 175      23   17.87500
## 176      24   17.87500
## 177      14   12.89375
## 178      18   12.89375
## 179      11   12.89375
## 180       0   12.89375
## 181      14   16.29375
## 182      11   16.29375
## 183      14   16.29375
## 184      23   16.29375
## 185      17   14.90625
## 186      25   14.90625
## 187      12   14.90625
## 188      17   14.90625
## 189      12   13.59375
## 190      14   13.59375
## 191       0   13.59375
## 192      22   13.59375
## 193      18   15.74792
## 194      18   15.74792
## 195      15   15.74792
## 196      20   15.74792
## 197      13   19.14792
## 198      10   19.14792
## 199      22   19.14792
## 200      19   19.14792
## 201      22   17.76042
## 202      20   17.76042
## 203      20   17.76042
## 204       0   17.76042
## 205      19   16.44792
```

```
## 206      19   16.44792
## 207      13   16.44792
## 208      16   16.44792
## 209      20   11.46667
## 210      15   11.46667
## 211      16   11.46667
## 212       0   11.46667
## 213      14   14.86667
## 214      22   14.86667
## 215      11   14.86667
## 216      10   14.86667
## 217      17   13.47917
## 218      21   13.47917
## 219      13   13.47917
## 220      22   13.47917
## 221       0   12.16667
## 222      15   12.16667
## 223       0   12.16667
## 224      22   12.16667
## 225      73   69.06042
## 226      73   69.06042
## 227      67   69.06042
## 228      69   69.06042
## 229      72   72.46042
## 230      69   72.46042
## 231      71   72.46042
## 232      73   72.46042
## 233      69   71.07292
## 234      71   71.07292
## 235      72   71.07292
## 236      70   71.07292
## 237      72   69.76042
## 238      73   69.76042
## 239      67   69.76042
## 240      67   69.76042
## 241      73   64.77917
## 242      72   64.77917
## 243      70   64.77917
## 244       0   64.77917
## 245      69   68.17917
## 246      73   68.17917
## 247      72   68.17917
## 248      67   68.17917
## 249      70   66.79167
## 250      70   66.79167
## 251      72   66.79167
## 252      69   66.79167
## 253      71   65.47917
## 254      70   65.47917
## 255       0   65.47917
```

```
## 256      69   65.47917
## 257      12   14.28958
## 258      15   14.28958
## 259      16   14.28958
## 260      11   14.28958
## 261      15   17.68958
## 262      19   17.68958
## 263      12   17.68958
## 264      17   17.68958
## 265      17   16.30208
## 266      20   16.30208
## 267      11   16.30208
## 268      10   16.30208
## 269      10   14.98958
## 270      10   14.98958
## 271      19   14.98958
## 272      17   14.98958
## 273      12   10.00833
## 274      18   10.00833
## 275      18   10.00833
## 276       0   10.00833
## 277      12   13.40833
## 278      18   13.40833
## 279      20   13.40833
## 280      11   13.40833
## 281      20   12.02083
## 282      10   12.02083
## 283      19   12.02083
## 284      21   12.02083
## 285      20   10.70833
## 286      10   10.70833
## 287       0   10.70833
## 288      19   10.70833
## 289      15   12.86250
## 290      13   12.86250
## 291      15   12.86250
## 292      12   12.86250
## 293      12   16.26250
## 294      22   16.26250
## 295      15   16.26250
## 296      21   16.26250
## 297      10   14.87500
## 298      20   14.87500
## 299      10   14.87500
## 300       0   14.87500
## 301      19   13.56250
## 302      22   13.56250
## 303      13   13.56250
## 304      14   13.56250
## 305      10    8.58125
```

```
## 306      20    8.58125
## 307      20    8.58125
## 308       0    8.58125
## 309      11   11.98125
## 310      16   11.98125
## 311      14   11.98125
## 312      22   11.98125
## 313      19   10.59375
## 314      18   10.59375
## 315      10   10.59375
## 316      19   10.59375
## 317       0    9.28125
## 318      13    9.28125
## 319       0    9.28125
## 320      21    9.28125
```