Assignment 4: Using statistical methods to examine factors that give rise to community engagement
(DUE FRIDAY, 12/3)

For each question, provide (1) the model used, (2) results (e.g., beta cofficients, F-statistics, p-value, graphs, etc.), and (3) the interpretation of the results.)

1. How does ethnic heterogeneity affect the poverty level for the given 28 cities, when controlling for the percentage of citizens?

   Using a linear regression:

```
Call:
lm(formula = model_1, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.131093 -0.065545 -0.005554  0.064753  0.193785

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.05859    0.76457  -0.077    0.940
ethnic_heterogeneity -0.59733    0.74048  -0.807    0.427
citizen_percent       0.47085    0.46563   1.011    0.322

Residual standard error: 0.08608 on 25 degrees of freedom
Multiple R-squared:  0.1213,      Adjusted R-squared:  0.05098
F-statistic: 1.725 on 2 and 25 DF,  p-value: 0.1987
```

   Ethnic heterogeneity, when controlling for the percentage of citizens, does not appear to affect the poverty level for the 28 cities. This is indicated by the high p-value for ethnic heterogeneity in the regression.

2. How does poverty affect people's participation in events (i.e., RSVPs) for the given 28 cities, when controlling for the population?

   Using a linear regression:

```
Call:
lm(formula = model_1, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-21056 -12902  -1198   6383  43173

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.448e+03  4.010e+03   1.858    0.075 .
poverty_index -1.787e+05  3.513e+04  -5.087 2.97e-05 ***
pop            1.624e-02  1.910e-03   8.505 7.56e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15990 on 25 degrees of freedom
Multiple R-squared:  0.7797,      Adjusted R-squared:  0.7621
F-statistic: 44.24 on 2 and 25 DF,  p-value: 6.137e-09
```

   The low p-value for poverty_index, when controlling for population, suggest that it has a significant effect on people’s participation in events.

3. How does poverty affect people's participation in events (i.e., RSVPs) for the given 28 cities, when controlling for the population and Gini index?

   Using a linear regression:

```
Call:
lm(formula = model_1, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-22068 -13005   -516   5766  41638

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.266e+04  9.184e+04  -0.247    0.807
poverty_index -1.828e+05  3.786e+04  -4.828 6.43e-05 ***
pop            1.587e-02  2.249e-03   7.057 2.69e-07 ***
gini           6.464e+04  1.970e+05   0.328    0.746
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16290 on 24 degrees of freedom
Multiple R-squared:  0.7807,      Adjusted R-squared:  0.7533
F-statistic: 28.48 on 3 and 24 DF,  p-value: 4.459e-08
```

The low p-values in the regression suggest that poverty has a strong effect on people's participation in events (i.e., RSVPs) for the given 28 cities, when controlling for the population and Gini index.

4. How is socio-economic inequality (i.e., Gini index) related to poverty for the given 28 cities?

   Using a linear regression:

```
Call:
lm(formula = model_1, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.040667 -0.012071 -0.003312  0.013693  0.042985

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.473732   0.003544 133.666   <2e-16 ***
poverty_index 0.076558   0.040844   1.874   0.0721 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01875 on 26 degrees of freedom
Multiple R-squared:  0.119,       Adjusted R-squared:  0.08516
F-statistic: 3.513 on 1 and 26 DF,  p-value: 0.07215
```

   Socio-economic inequality (i.e., Gini index) appears to be strongly correlated to poverty for the given 28 cities. This is evident in the results of the multi-level regression.

5. How is socio-economic inequality (i.e., Gini index) related to the number of events per capita for the given 28 cities?

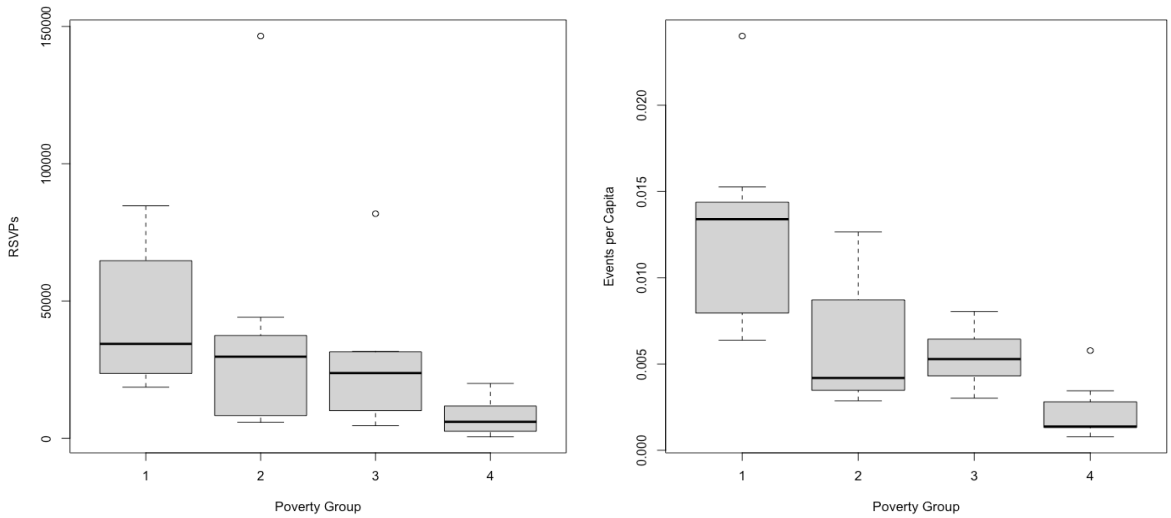
   Using a multi-level regression:

```
boundary (singular) fit: see ?isSingular
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: model_1
   Data: data

REML criterion at convergence: -135.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5951 -0.6722 -0.2654  0.7186  2.0877
```

```
Random effects:
 Groups    Name        Variance  Std.Dev.
 state     (Intercept) 0.0000000 0.00000
 Residual              0.0003731 0.01931
Number of obs: 28, groups:  state, 17

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  0.480048   0.005932 26.000000  80.931   <2e-16 ***
std_events  -0.947886   0.701747 26.000000  -1.351    0.188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr)
std_events -0.788
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see ?isSingular
```

Socio-economic inequality (i.e., Gini index) is negatively correlated to the number of events per capita for the given 28 cities. This is evident in the linear regression calculation.

6. When 28 cities are categorized into 4 groups based on poverty level (7 cities in each group, based on the order of poverty level), are there systematic differences in their RSVPs and the number of events per capita between different groups of poverty? Examine this question using ANOVA. Also, provide a Box plot for showing the differences between the groups. ANOVA was not covered in the class, but you can use the "anova()" function instead of "summary()" to see the significances in R.



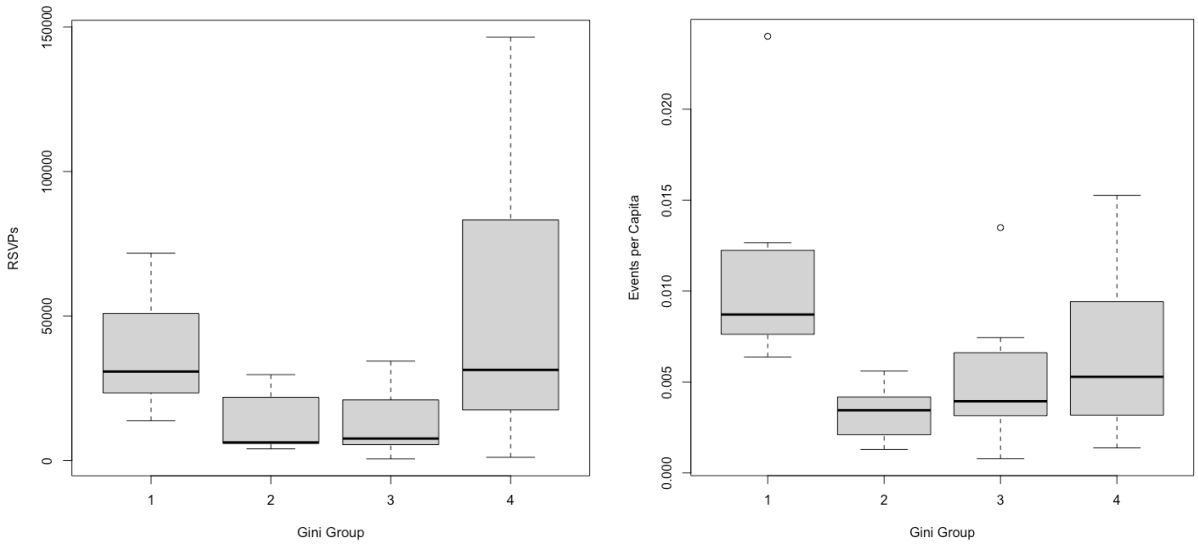
Using the anova() function:

```
Analysis of Variance Table

Response: rsvp
              Df     Sum Sq    Mean Sq F value  Pr(>F)
poverty_group  1 5.2546e+09 5254588774   5.746 0.02401 *
Residuals     26 2.3776e+10  914480287
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Analysis of Variance Table

Response: std_events
```

```
               Df     Sum Sq    Mean Sq F value    Pr(>F)
poverty_group   1 0.00035532 0.00035532  22.967 5.813e-05 ***
Residuals      26 0.00040223 0.00001547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the boxplots, there are clear differences among the groups. Further, the large F value (and resulting small P value) suggests a strong relationship for both variables.

7. When 28 cities are categorized into 4 groups based on the level of socio-economic inequality, i.e., Gini index (7 cities in each group), are there any systematic differences in their RSVPs and the number of events per capita? Examine this question using ANOVA. Also, provide a Box plot for showing the differences between the groups.



Using the anova() function:

```
Analysis of Variance Table

Response: rsvp
           Df     Sum Sq    Mean Sq F value Pr(>F)
gini_group  1 8.6002e+08  860024629  0.7937 0.3811
Residuals  26 2.8171e+10 1083501985


Analysis of Variance Table

Response: std_events
           Df     Sum Sq    Mean Sq F value Pr(>F)
gini_group  1 0.00004606 4.6058e-05  1.6831 0.2059
Residuals  26 0.00071149 2.7365e-05
```

The boxplots for RSVP and events per capita are loosely similar. However, the F and P values do not suggest a strong relationship to the Gini groups.