Name: Your name here
Due: 2024/11/04

Homework 7

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, November 11th at 11:59 pm. The purpose of this assignment is to practice using time series regression methods to analyze data.

Boreal forest grouse The capercaillie (*Tetrao urogallus*) and black grouse (*Tetrao tetrix*) are two species of grouse native to boreal forests in Norway. Recently, a study[1] was conducted to better understand the temporal dynamics of breeding success among these two species of grouse, which is defined as the ratio of the number of chicks born to a community and the number of hens within that community. Researchers leveraged a unique data set that tracked the breeding success of two populations of grouse over 41 years, and tracked a number of relevant variables, including snow depth and measures of predation.

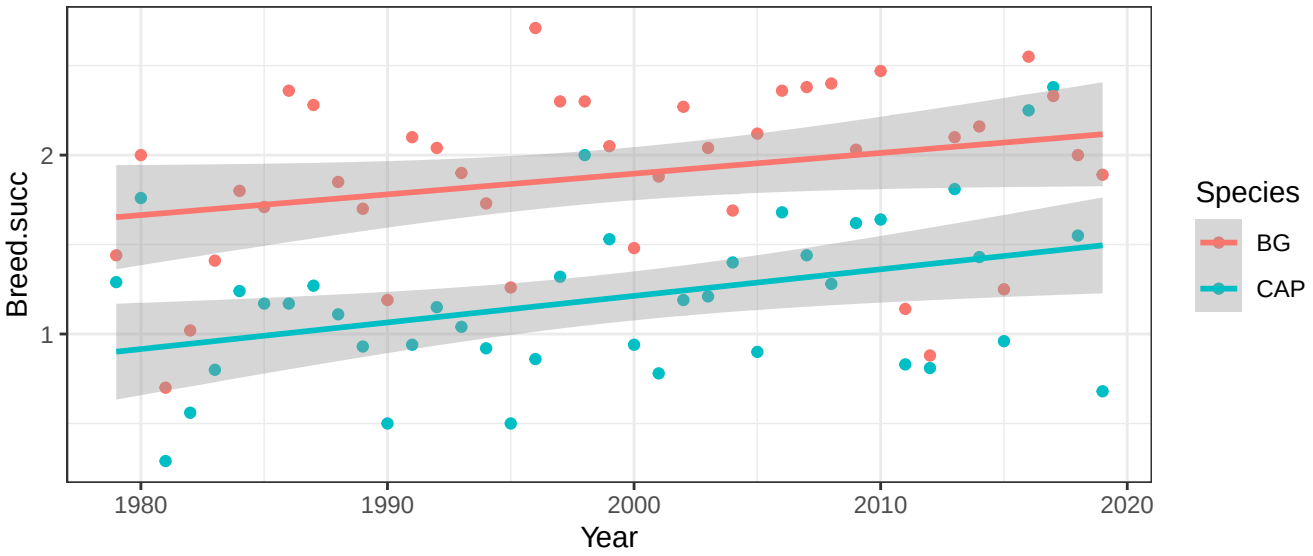
Black grouse (left) and capercaillie (right)

[1] Link to paper describing the study.

Part I [14 pt] Suppose that we are interested in determining whether there is a difference in breeding success between the two species after accounting for changes over time.

1. [2 pt] Create a plot that visualizes the research question and comment on what the plot suggests about the research question.

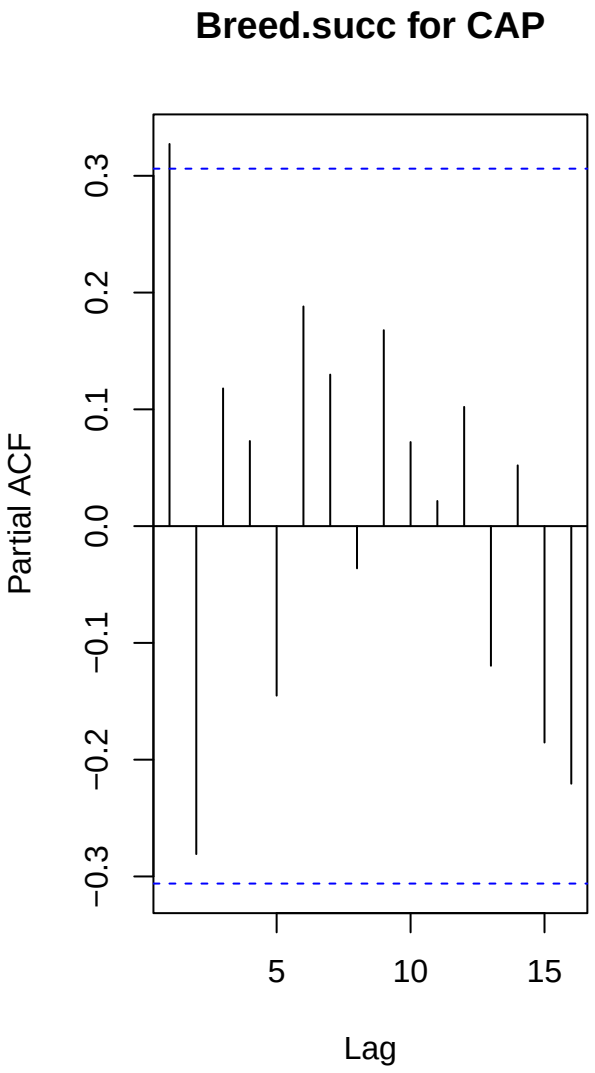
```
grouse <- read_csv("grouse.csv")
grouse %>%
  ggplot() +
  geom_point(aes(x = Year, y = Breed.succ, col = Species)) +
  geom_smooth(aes(x = Year, y = Breed.succ, col = Species), method = "lm") +
  theme_bw()
```

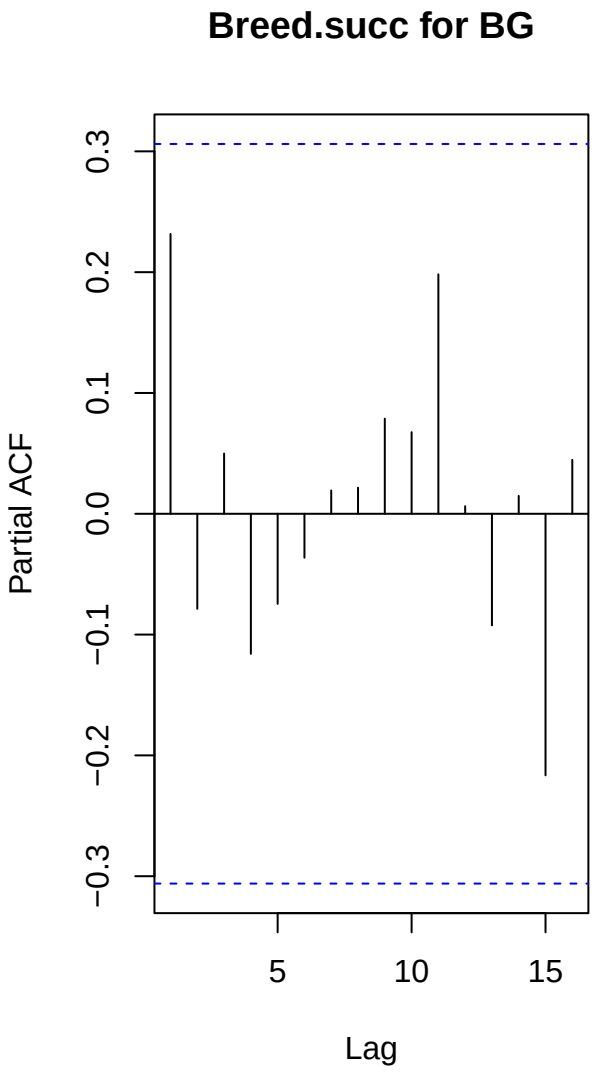


Based on the smoothers, there does appear to be a difference in breeding success after accounting for time. On average, the black grouse seem to have a higher breeding success than the capercaillie, after accounting for time.

2. [3 pt] Before fitting any models, create a pacf plot of the raw `Breed.succ` for both species and comment on what the plot suggests about autocorrelation.

```
par(mfrow = c(1,2))
pacf(grouse$Breed.succ[1:41], main = "Breed.succ for CAP")
pacf(grouse$Breed.succ[42:82], main = "Breed.succ for BG")
```





Before accounting for any other variables, there is some evidence of autocorrelation in the `Breed.succ` for both species. While neither species has many spikes that cross the "significant" threshold, the magnitude of those correlations are fairly high. There is somewhat convincing evidence of an AR(2) process for the capercaillie species, as there are fairly large spikes at lags 1 and 2 for that species.

3. [1 pt] Fit an ordinary least squares model to address the research question.

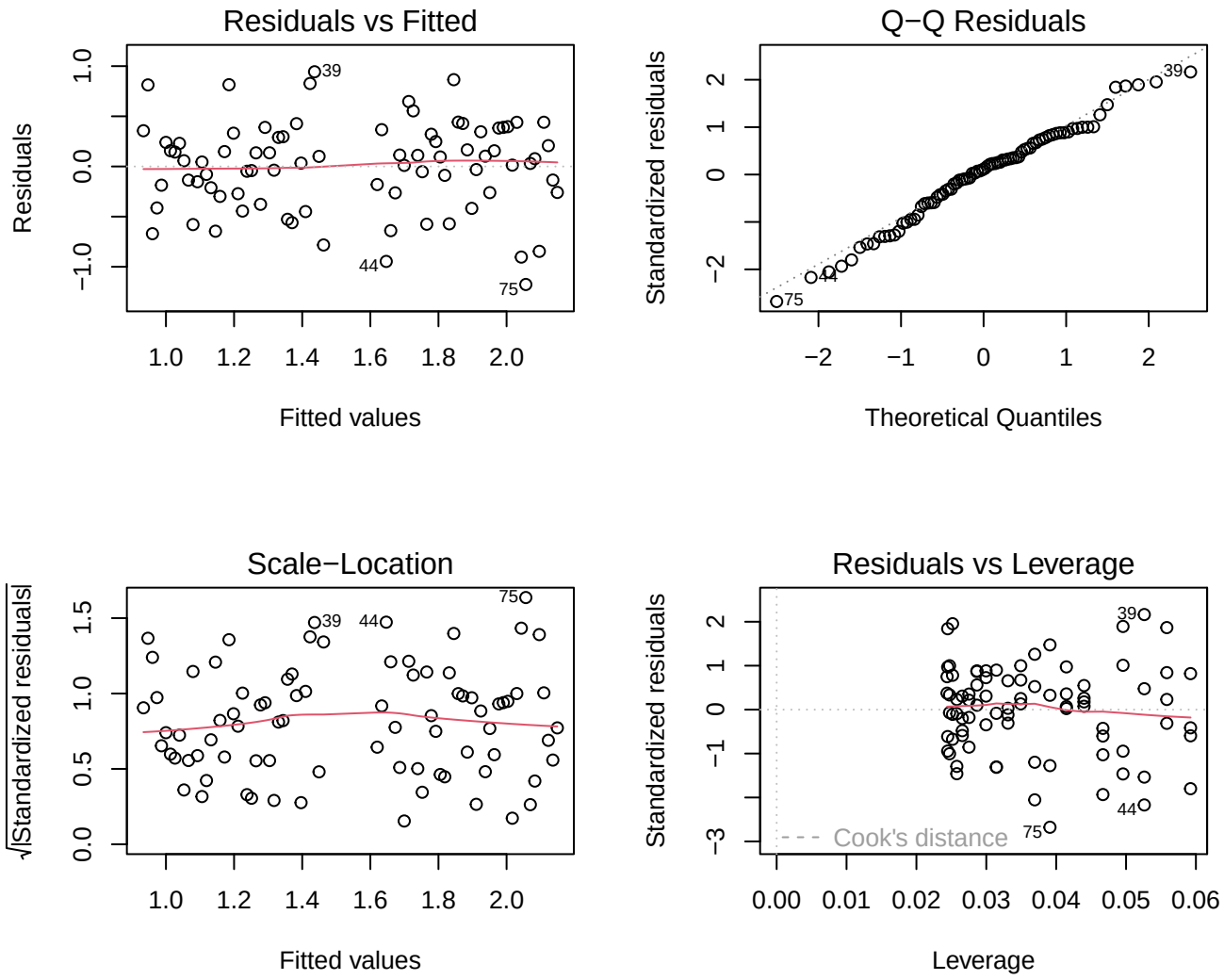
```
ols_fit <- lm(Breed.succ ~ Year + Species, grouse)
summary(ols_fit)
```

```
Call:
lm(formula = Breed.succ ~ Year + Species, data = grouse)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.17649 -0.26252  0.05068  0.31491  0.94376

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.541171   8.362527  -2.935  0.00437 **
Year          0.013220   0.004183   3.160  0.00224 **
SpeciesCAP   -0.686341   0.098993  -6.933    1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4482 on 79 degrees of freedom
Multiple R-squared:  0.4236, Adjusted R-squared:  0.409
F-statistic: 29.03 on 2 and 79 DF,  p-value: 3.538e-10
```

4. [3 pt] Assess the linearity, normality, and constant variance assumptions of the ordinary least squares model.

```r
par(mfrow = c(2,2))
plot(ols_fit)
```

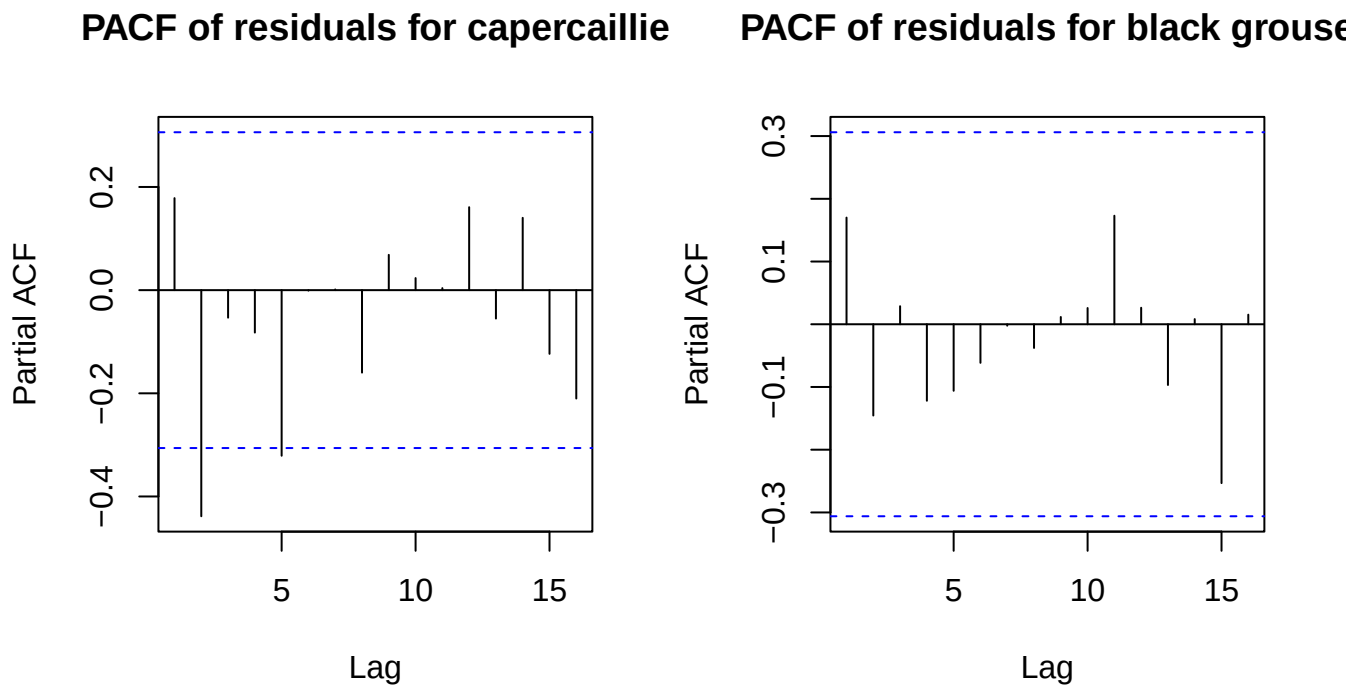


Normality: The points follow the hypothesized quantiles in the QQ plot quite well, suggesting that the distribution of the residuals is consistent with the normality assumption.
Linearity: There does not appear to be any obvious leftover trends in the residuals vs. fitted values plot, suggesting that the linearity assumption is reasonably satisfied.
Constant variance: Finally, we do not see any evidence of fanning or unequal variances in the residuals vs fitted plot, nor do we see any trends in the scale-location plot, suggesting that the constant variance assumption is reasonably satisfied.

5. [3 pt] Assess the independence assumption for the model, referencing plots as needed. (hint: look at a pacf plot!)

> **!** Important
>
> When creating ACF and PACF plots of the residuals, the observations must be in sequential order!

```
par(mfrow = c(1, 2))
pacf(resid(ols_fit)[1:41], main = "PACF of residuals for capercaillie")
pacf(resid(ols_fit)[42:82], main = "PACF of residuals for black grouse")
```



In general, we would expect the independence assumption to be violated, as we measured these populations of birds over time, and would therefore expect measurements taken close together in time to be similar. When looking at the PACF plot, there is fairly strong evidence of an AR(2) process in the residuals for the capercaillie species.

6. [4 pt] Regardless of your previous assessments, answer the research question, supporting your answer with an appropriate statistical test and confidence interval for parameter of interest.

```
confint(ols_fit)
```

```
                   2.5 %      97.5 %
(Intercept) -41.186364983 -7.89597648
Year          0.004893043  0.02154598
```

```
SpeciesCAP   -0.883381678 -0.48930125
```

After accounting for year, there is strong evidence to suggest that the average breeding success rate depends on the species, as the estimated effect of species after accounting for year is $\widehat{\beta}_{\text{speciesCap}} = -0.686$ with a standard error of 0.099, resulting in a t-statistic of -6.933 with 79 degrees of freedom, yielding a p-value of less than 0.0001. We are 95% confident that, for a particular year, the mean breeding rate for capercaillie grouse is between .489 and .883 less than the mean breeding rate for black grouse.

Part II [10 pt] In this part, we specifically focus on the error structure.

7. [2 pt] Is the above ordinary least squares model sufficient for this analysis? Or do we require a more complicated time series model? Explain your answer.

> The PACF plot suggests there may be some evidence of an AR(2) process at work here, and it is probably safer to go ahead and account for it.

8. [2 pt] Regardless of your answer to the previous question, refit the regression model imposing an AR(2) process on the residuals using generalized least squares and print a summary of the model.

> **Tip**
>
> By default, the `corARMA` function assumes that each observation represents a time point, and that your observations are ordered according to time. In order to account for the time structure between the groups, you will need to use the following correlation structure: `corARMA(form = ~ 1 | Species, p = 2, q = 0)`, which results in a pooled estimate of the autoregressive parameters between the two species.

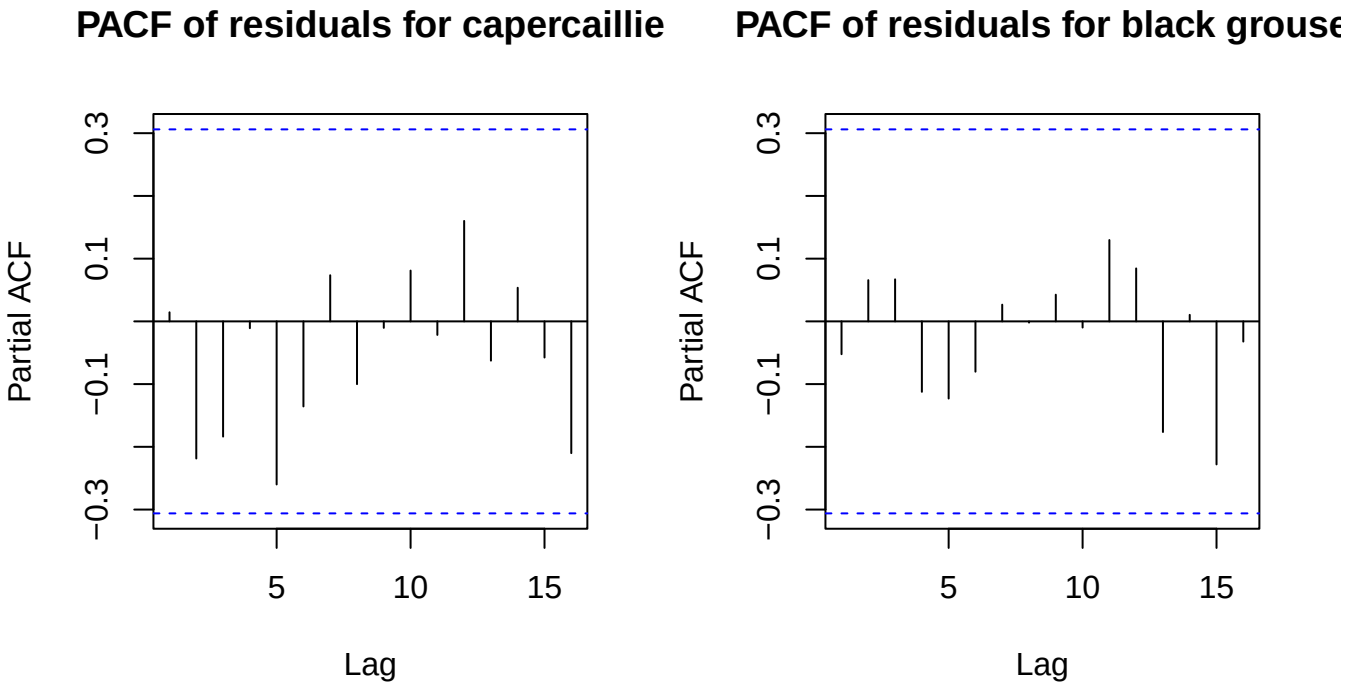
```
gls_fit <- gls(Breed.succ ~ Year + Species, grouse, correlation = corARMA(form = ~ 1
summary(gls_fit)
```

```
Generalized least squares fit by REML
  Model: Breed.succ ~ Year + Species
  Data: grouse
       AIC      BIC   logLik
  116.9436 131.1603 -52.4718

Correlation Structure: ARMA(2,0)
 Formula: ~1 | Species
 Parameter estimate(s):
      Phi1       Phi2
 0.2592092 -0.2707312

Coefficients:
                 Value Std.Error   t-value p-value
(Intercept) -25.876020  7.940278 -3.258831  0.0017
Year          0.013887  0.003972  3.496124  0.0008
SpeciesCAP   -0.690863  0.092718 -7.451250  0.0000

 Correlation:
           (Intr) Year
Year       -1.000
SpeciesCAP -0.006  0.000
```

```
Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.6452402 -0.5929702  0.1377739  0.7142194  2.0953409

Residual standard error: 0.4475043
Degrees of freedom: 82 total; 79 residual
```

9. [2 pt] Reassess the independence assumption using the residuals from the gls fit.

```
par(mfrow = c(1, 2))
pacf(resid(gls_fit, "normalized")[1:41], main = "PACF of residuals for capercaillie")
pacf(resid(gls_fit, "normalized")[42:82], main = "PACF of residuals for black grouse")
```



It looks much better. :) For both species, there are no meaningful spikes in the PACF plots.

10. [4 pt] Compare and contrast the p-value and confidence interval obtained from the gls fit to those obtained from the ols fit. Provide an explanation for what you notice.

```
confint(gls_fit)
```

```
                   2.5 %       97.5 %
(Intercept) -41.438679211 -10.31336173
Year          0.006101612   0.02167153
```

```
SpeciesCAP   -0.872586516  -0.50913964
```

In this case, the confidence intervals are actually *narrower* and the p-values are *smaller* for the year and species effects. We end up with narrower intervals and smaller p-values because the autocorrelation in the residuals is *negative*, which means our original estimates of the standard errors were actually inflated.

Part III [1 pt] List at least one and up to three other people you will be working with for the project (i.e. between two and four people per group).