Question 1

Below I will describe a number of different study designs. You will be asked to

(1) Identify the null hypothesis
(2) Identify the correct statistical test for this hypothesis

Part A: For this study, I am interested in determining if a student's major (Humanities/STEM/Social Sciences) is associated with their final exam score in STA-209.

Part B: A two-day workshop for learning basic R has been created, where attendees are tested in their R skills both prior to the workshop and after the workshop has been completed. For each of these tests, a numeric score is given. We wish to determine whether or not the workshop has been effective in improving the R skill of the attendees.

Part C: Binge drinking is defined as a pattern of drinking that involves consuming 5 or more standard drinks within 2 hours. Respondents of a survey were asked for their sex and whether or not they have engaged in binge drinking more than twice in the previous week. We wish to determine whether or not there is a difference in binge drinking patterns between men and women.

Question 2

Cocaine addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to use cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo (lithium is the standard treatment for cocaine addiction). One third of the subjects were randomly assigned to each treatment group with the following results:

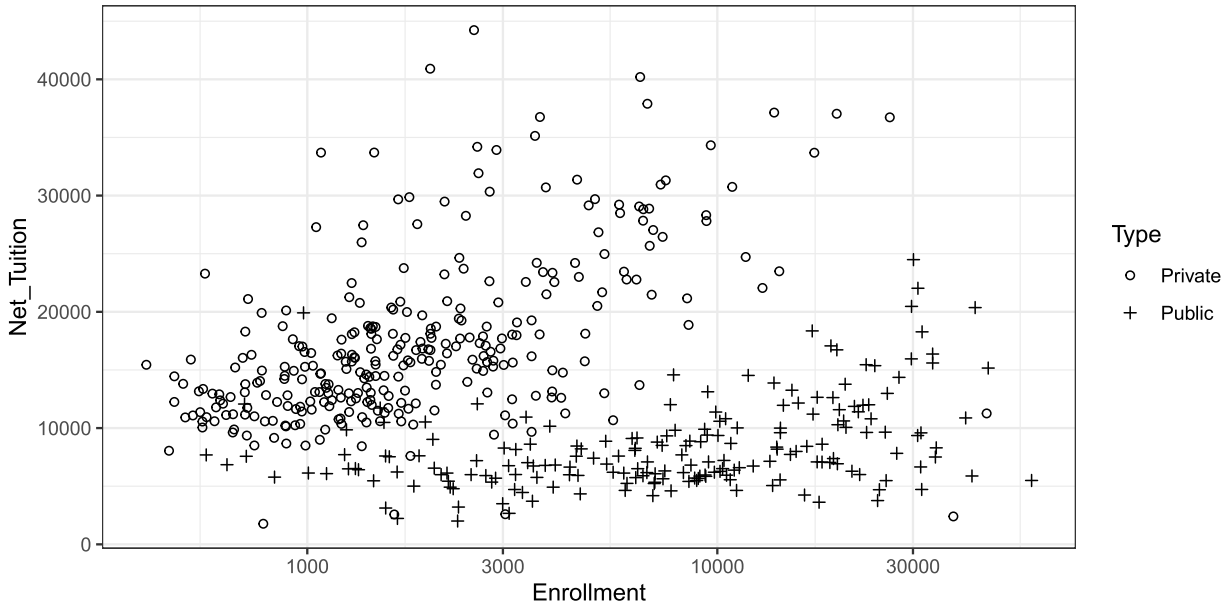
	Relapse	No Relapse
Desipramine	10	14
Lithium	18	6
Placebo	20	4

Part A: What type of plot would you use to visually display these results

Part B: Describe the null hypothesis of this study and construct a table of expected counts under the assumption of the null hypothesis.

Part C: Compute the χ^2 statistic for the null hypothesis in B. Based on this, what conclusion would you reach if you were testing at the $\alpha = 0.05$ level?

Question 3



Model 1:

```
lm(formula = Net_Tuition ~ Enrollment, data = college)

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 14225.3137   272.8034    52.1 <0.0000000000000002 ***
Enrollment     -0.0820     0.0265    -3.1               0.002 **

Residual standard error: 7180 on 1093 degrees of freedom
Multiple R-squared:  0.00869,   Adjusted R-squared:  0.00779
F-statistic: 9.58 on 1 and 1093 DF,  p-value: 0.00201
```

Model 2:

```
lm(formula = Net_Tuition ~ Enrollment + Type, data = college)

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  5746.1019   377.2481    15.2 <0.0000000000000002 ***
Enrollment      0.2533     0.0239    10.6 <0.0000000000000002 ***
TypePrivate 10808.5970   398.6370    27.1 <0.0000000000000002 ***

Residual standard error: 5550 on 1092 degrees of freedom
Multiple R-squared:  0.408, Adjusted R-squared:  0.406
F-statistic:  376 on 2 and 1092 DF,  p-value: <0.0000000000000002
```

Part A: For this part, consider **Model 1** from above. What is the null hypothesis in linear regression? Based on the summary output, how would you describe the relationship between enrollment and tuition?

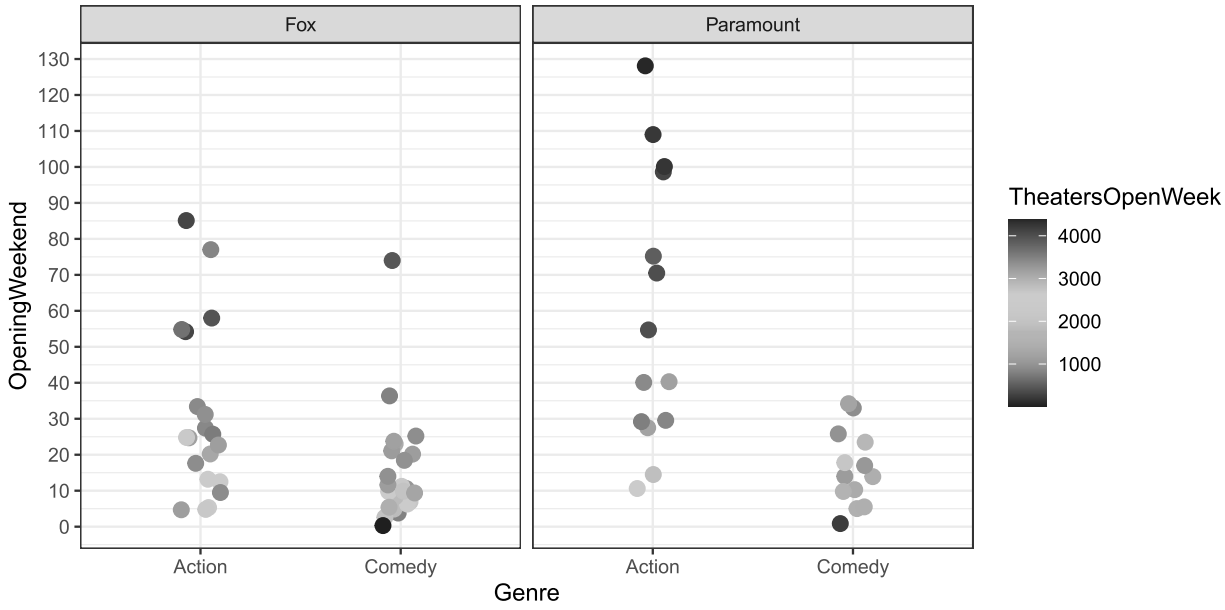
Part B: Now consider **Model 2**, which includes an indicator for whether or not a college is private. How would you interpret the intercept in this model? Is this a meaningful value in this model?

Part C: Compare the coefficient for Enrollment between **Model 1** and **Model 2**. What has changed? In other words, what impact has adding an indicator for Private had on this value, and why did it result in such a drastic change?

Question 4

(This page can be ripped off)

Included below are data from 70 Hollywood films released between 2007 and 2001. Movies in this dataset include Action and Comedy films from two major studios, Fox and Paramount. The plot below illustrates the total sales over a film's opening weekend, with a color aesthetic to indicate the number of theaters in which the film was shown: dark red corresponds to a film showing in a large number of theaters, while dark blue indicates that it was shown in relatively few theaters.



Below is summary information for a linear regression model with revenue from the opening weekend (`OpeningWeekend`) serving as the *dependent variable* and with film studio (`LeadStudio`) and genre (`Genre`) serving as the *independent variables*.

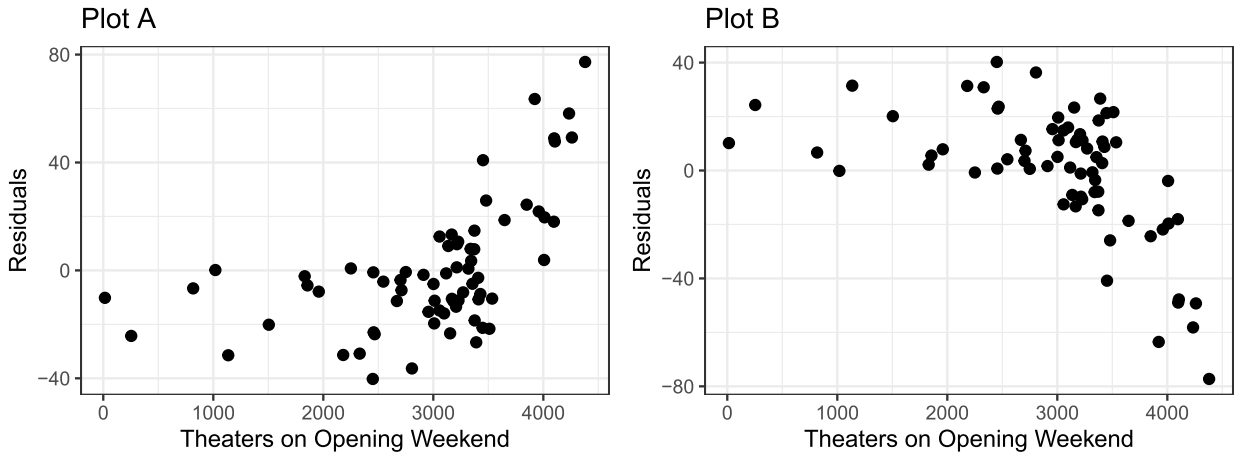
```
> lm(OpeningWeekend ~ Genre + LeadStudio, movies) %>% summary()

Coefficients:
                   Estimate Std. Error t value      Pr(>|t|)
(Intercept)           36.15       4.81    7.52 0.00000000018 ***
GenreComedy          -25.69       5.78   -4.44 0.00003427689 ***
LeadStudioParamount   14.69       5.94    2.47         0.016 *

Residual standard error: 24.2 on 67 degrees of freedom
Multiple R-squared:  0.288, Adjusted R-squared:  0.266
F-statistic: 13.5 on 2 and 67 DF,  p-value: 0.0000116
```

You will use these plots and summary data to answer the following questions:

Part A: Give an interpretation of the intercept of this model. Is this meaningful?

Part B: Again using the summary information, find the predicted opening weekend revenue for each genre/studio combination (i.e., predicted opening revenue for a Comedy film from Fox). With these predictions, *mark horizontal lines on the plot above indicating the predicted opening weekend revenue for each category.*

Part C: We are now interested in determining if the variable for the number of theaters showing a film on opening weekend (`TheatersOpenWeek`) should be included in our model. We will do this by plotting the residuals of the model above against the missing variables. Using the changes you made to the plot above in Part B, determine which of the plots below shows the correct association between the model residuals and the number of theaters on opening weekend (circle one). Include 1-2 sentences to justify your answer.



Part D: Below is the updated model for predicting revenue on opening weekend, now including the variable for the number of theaters:

```
> lm(OpeningWeekend ~ Genre + LeadStudio + TheatersOpenWeek, movies) %>% summary()


Coefficients:
                     Estimate Std. Error t value    Pr(>|t|)
(Intercept)         -21.48761    9.95081   -2.16      0.0345 *
GenreComedy         -13.69726    4.99129   -2.74      0.0078 **
LeadStudioParamount   8.74991    4.82823    1.81      0.0745 .
TheatersOpenWeek      0.01804    0.00287    6.28 0.000000031 ***

Residual standard error: 19.3 on 66 degrees of freedom
Multiple R-squared:  0.554, Adjusted R-squared:  0.534
F-statistic: 27.3 on 3 and 66 DF,  p-value: 0.0000000000133
```

Given an interpretation of the intercept in this model. Is this meaningful?

Part E: Consider the presented linear models, both with and without the variable `TheatersOpenWeek`. Based on the summary information, which would you prefer to use to predict revenue on opening weekend? Briefly justify your answer.