Instructions for the final movie
Dr. Jain wants to see the following in the video submission describing our research project.

- Hypothesis
- How you designed the experiment
- How you conducted an experiment
- How you analyzed the data
- Draw conclusions from the data (either the hypothesis was correct or inconclusive)
- Discuss your conclusions (especially with respect to the paper you based your experiment on)

Also include information about

- Examples of stimuli
- Outliers, and how you handled them
- What statistical tests you use
- Why do you think your conclusions are sound

The video should be a maximum of five minutes and does not need to include voice narration. The video should be uploaded to YouTube, and the link shared with the rest of the class (details about sharing forthcoming).

Linear Regression
Some research questions deal with determining whether multiple groups differ from each other (e.g. do people perform better on exams when they have a full night of sleep compared to no sleep). Other research questions try and develop predictive models (e.g. can exam performance be predicted based on the amount of sleep you got the previous night). This second example involves fitting a model to data, which can be accomplished using linear regression.

Predictive models predict the value of a dependent variable (e.g. exam performance) based on the given value of the independent variable (e.g. hours of sleep). This can be described as \( Y = f(X) \), where \( Y \) is the dependent variable, \( X \) is the independent variable, and \( f() \) is the predictive model. Once a predictive model has been learned, you can generalize the results and predict responses for independent values you have never seen before.

Another example is the relationship between the age of a child (\( X \)) and the height of a child (\( Y \)). This relationship could look as follows.
This figure describes an exponential model. The figure shown below describes a linear model, which we are discussing today.

The below graph contains 7 data points. Many different lines can be drawn to fit this data (3 are shown). While any of these lines could explain the data, we want to find the line that *best* explains the data.

What is the best explanation? How do you find it?

Linear regressions are presented as follows \( y_i: B_0 + B_1 * X_i + \epsilon_i \)
Where \( y_i \) equals the predicted value of the dependent variable given the independent variable equals \( X_i \). \( B_0 \) and \( B_1 \) are the two constants that describe the linear model, and \( \epsilon_i \) is the error term describing the error in the model and in the measurements.

It is important to distinguish between \( y_i \) and \( Y_i \), where \( Y_i \) is an actual measurement and is the \( y_i \) predicted value of a given \( X_i \). \( y_i - Y_i \) is the residual, the difference between a predicted value and a measured value. Residuals are illustrated in the figure below as vertical lines connecting actual data points to the predicted values.

The most widely accepted way of finding the line of best fit is to find the line that minimizes the residuals. You find this line using the method of least squares. The goal of this method is to minimize the squared difference in the residuals. \( SS_R = \sum (Y_i - y_i)^2 \)

One of the most naïve models describing data is to always predict the mean value of the observed data. This would give a model like the one shown below.

This is generally one of the worst models you can use. When assessing a specific linear model, we compare it to the sum of squares of the residuals of this mean model, given by \( SS_T = \sum (Y_i - \bar{Y})^2 \). The model sum of squares is given by \( SS_M = SS_T - SS_R \). We can then calculate the goodness of fit of a specific linear model using the following equation: \( R^2 = SS_M / SS_T \). This is also called the coefficient of determination.

**Demo in R**

R includes several built-in data sets. This example uses the data set *faithful*, which contains measurements from the geyser Old Faithful. Two measurements are included in the dataset *faithful*: the
duration of an eruption and the time since the last eruption. We want to ask the question is there a linear relationship between the duration of an eruption and the time since the last eruption. In this case, time since last eruption is our independent variable and duration is our dependent variable.

We can first take a look at the data using the command shown below. Head allows you to preview a dataset.

```
head(faithful)
```

Next we can fit a linear model using the `lm` command. The model is defined using the syntax `dependent variable ~ independent variable`.

```
eruption.lm = lm(eruptions ~ waiting, data=faithful)
```

We can then use the `coefficients` command to fetch the coefficients of the linear model we just built.

```
coeffs = coefficients(eruption.lm)
```

We can then calculate the predicted eruption duration given the time since the last eruption. In this case, the time since last eruption is set to 80.

```
w=80
d=coeffs[1]+coeffs[2]*w
d
```

Once you have the model, you don't need to hang on to the data itself. All you need to do is store the model. This makes predictive models very powerful and very easy to use. Linear models are among the most basic types of models. Many other models exists.

You can use the command `summary(eruption.lm)` to gain additional information about the model. Two values are particularly important. R-squared tells you the amount of variance accounted for in the data. In this case, our model accounts for 81.08% of the variance in the data. We can also do a hypothesis test to see if the slope of the line is different from zero. If the p-value is less than 0.05, then there is a statistically significant relationship between X and Y. In this case, the p-value for the slope is p < 2^{-16}, which is highly significant.

Finally, we can plot the linear model against the data using the following two commands. The first command plots the data and the second command draws the linear model. The final plot is shown below the commands.

```
plot(eruptions~waiting, data=faithful)
abline(coeffs)
```
Additional Example

R includes an additional dataset called ChickWeight. This dataset includes measurements of the weight of fifty different chickens over time. These chickens were fed one of four data sets. We want to build linear models for each of these four diets (note: in a real analysis, we would want to do this analysis using multiple-linear regression to explore the effect of the four diets simultaneously).

First we first need to subset the data based on diet. This will create four new datasets that only contain the chickens who were fed a specific diet. We can do this as follows

```r
diet1 <- subset(ChickWeight, ChickWeight$Diet==1)
diet2 <- subset(ChickWeight, ChickWeight$Diet==2)
diet3 <- subset(ChickWeight, ChickWeight$Diet==3)
diet4 <- subset(ChickWeight, ChickWeight$Diet==4)
```

We can then build linear models for each of these new datasets using the following four commands.

```r
diet1.lm <- lm(weight ~ Time, data=diet1)
diet2.lm <- lm(weight ~ Time, data=diet2)
diet3.lm <- lm(weight ~ Time, data=diet3)
diet4.lm <- lm(weight ~ Time, data=diet4)
```

Next we’ll calculate the coefficients for these four models

```r
diet1.coefs <- coefficients(diet1.lm)
diet2.coefs <- coefficients(diet2.lm)
diet3.coefs <- coefficients(diet3.lm)
diet4.coefs <- coefficients(diet4.lm)
```

The calculate coefficients are shown in the table below. Of these four models, it appears that diet 3 promotes the fastest weight gain, as this is the model with the highest slope. Diet 1 appears to be the worst diet (assuming weight gain is the goal), as it has the smallest slope.

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet1</td>
<td>30.930</td>
<td>6.841</td>
</tr>
<tr>
<td>Diet2</td>
<td>28.633</td>
<td>8.609</td>
</tr>
<tr>
<td>Diet3</td>
<td>18.250</td>
<td>11.423</td>
</tr>
<tr>
<td>Diet4</td>
<td>30.792</td>
<td>9.714</td>
</tr>
</tbody>
</table>

The R-squared values and p-values are shown in the table below. All of these models show a statistically significant relationship between time and weight. They also all explain a good deal of the variance in the data.

<table>
<thead>
<tr>
<th></th>
<th>R-squared</th>
<th>Slope p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet1</td>
<td>0.664</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Diet2</td>
<td>0.666</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Diet3</td>
<td>0.805</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Diet4</td>
<td>0.905</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>
Plots of the four linear models are shown below.