The dataset that I used came from a study on divergent thinking and mind wandering. Specifically, my goal was to determine if there were common responses on the creative thinking task, in which participants had to give a creative answer for the use of a knife and a brick. Then, I wanted to see if the commonality of those responses correlated with the interviewer rating of creativity. I also performed a similar linear model that they did, looking at the rate of task unrelated thoughts during the sustained attention response task, n back task, number stroop flanker task, and arrow flanker task. All the variables were numeric, except for the individual response variable and commonality of answers column that I inserted myself (categorical/strings).

First, I read in my 3 csv files using the read.csv function. Next, I renamed the participant ID columns in all 3 datasets separately, using the colnames function and indexing each dataset to the first column to rename it. I set each dataset's first column's new name to be ID, so that I could easily merge the datasets later on. Then, I had to switch the creative task datasets to wide format, as I wanted only one response for each variable for each unique participant. My ECDTdata was in wide format already, so I didn't need to worry about reshaping it. However, first I had to combine each Problem1RESP entry into a single entry. I did so using a for loop that ran through each unique ID (using the unique function) in the datasets and combined the Problem1RESP for each instance of that ID using the paste function. I then assigned the combined responses to a new column, called Response. Finally, I used the dcast function in the reshape2 package to convert CreativeBrickTask and CreativeKnifeTask into wide format. I also excluded all other columns other than the ID and newly combined AllResponses column because the data for each interviewer rating was already averaged and included in the ECDTdata dataframe, which I would merge with anyway.

After the reshaping, I merged the wide dataframes of the creative brick task and creative knife task using the merge function by ID. I also used the suffixes argument to differentiate between the 2 AllResponses columns. Lastly, I performed a second merge of the newly created CreativeTasks dataset and the ECDTdata by ID and saved that to be named allECDTdata. Then, I wrote allECDTdata to a csv file to look at before I did more data cleaning. First, I made the AllResponses columns uniform by using the tolower function. Then I removed all unnecessary spaces before the commas by using gsub and the regular expression "\\s*," (one white space 0 or more times followed by a comma) and replacing it with just a comma, reassigning it to the corresponding columns. Next, I averaged the creative rating columns together by ID to create new rated creative response frequency columns for the knife task and brick task. I used the apply function on all ECDTdata, selected all the rater columns, put 1 as the argument to combine them by row, and applied the mean function. Then, I performed a summary using the summary function to look at the range of ratings, which was -.63 to 4.47 for the brick task and -.5 to 3.28 for the knife task. It seems that the knife task responses had less creative responses overall than the brick task. I then removed all the individual rater scores from the data set using the subset function and -c() because they were no longer necessary after averaging the ratings.

Next, I scanned the responses for each task and tried to find some common answers. I then used grepl with the regex "draw|drawing" to isolate that common brick response and used the sum function to get a count, which was 30. Similarly, I also used grepl and sum to find the count of "break" in the brick responses, which was 78. I then searched for two common response in the creative knife responses and used grepl and the regex "\\bcarv[a-z]{1:3}" to capture a full word that started with carv and ended with 1-3 more letters. This pattern detected instances of carve, carving, and carver–the sum was 67. Lastly, I grepled for "cut" and the sum was 210.

Using these patterns, I made 2 new columns: CommonBrick and CommonKnife to indicate if a person had any of the common responses for the brick or knife task. I used an ifelse statement to do this, with the condition drawPattern|breakPattern (the grepl statement from earlier) for the brick column and carvePattern|cutPattern for the knife column, setting the entry to “common” or “uncommon”. If a participant had either response, they were marked with “common”, as having a common response. I then used the ggplot2 package to create barplots for both the Knife and Brick Common columns. I used the aes argument to set the x axis to be the column responses of common and uncommon, also using CommonBrick/Knife in the fill argument. Within the added geom_bar function, I set the position to be dodge so as not to stack the common/uncommon counts and then used scale_fill_manual to set the colors to be blue/purple and light blue/lavender for uncommon and common respectively. Lastly, I used the labs function to change the title, x axis, and y axis labels. I then saved the resulting graphs to a pdf document. Finally, I used the subset function to get rid of empty columns as well as all the test result columns, as I won’t be looking at them in my analysis. I then renamed the mind wandering rate columns using colnames and bracket notation. Lastly, I used the str function to look at the makeup of my database and resaved the csv file. There are 436 observations (participants) of 11 variables, including ID, mind wandering rate of 4 tasks, brick task responses, knife task responses, and creative response frequencies and common responses of both the brick and knife tasks. The CommonBrick, CommonKnife, AllResponsesBrick, and AllResponsesKnife are characters but the rest of the variables are numeric.

I first made three linear models using the lm function, with creative response frequency as the dependent variable and CommonBrick/Knife as predictors for the first 2. The 3rd linear

model included both creative response frequency variables (for knife and brick) and all the mind wandering rate variables as its predictors. I then performed a summary of these linear models and saved it as a text file using the sink function. Checking for any violations for assumptions, I used the plot function first to look at the 3 linear models I made. Looking at the plots of each, the residuals vs. fitted for CreativeResponseFreqBrick~CommonBrick and CreativeRespFreqKnife~Common Knife have a wide spread with some points having very high residuals. The points are not randomly scattered, but split into two sides of the spectrum which indicates heteroscedasticity of variance. The qqplots for both follow the line pretty closely until the 2nd quantile so this may indicate it's not a normal distribution. Scale location sees a very similar split pattern to the residuals vs. fitted plot again indicates it doesn't have homogenous variances. Residuals vs. leverage show that there's a possibility of outliers, particularly 204, 345, and 425 in the knife model but not anything of concern in the brick model. Looking at the full linear model, the spread of the residuals vs. fitted plot looks promising, but the qq residuals have the same issue in the 2nd quantile, which indicates a normality violation. Scale-location indicates homogeneity of variance but the residuals vs. leverage graph shows a lot of points above Cook's distance, so these points will have an impact on the regression coefficients. The shapiro.test function output looked like this, so I would have to correct for the violation of normality to do any further analysis: `shapiro.test(allECDTdata$CreativeResponseFreqBrick)`

```
        Shapiro-Wilk normality test

data:  allECDTdata$CreativeResponseFreqBrick
W = 0.78272, p-value < 2.2e-16
```

```
> shapiro.test(allECDTdata$CreativeResponseFreqKnife)
```

```
        Shapiro-Wilk normality test

data:  allECDTdata$CreativeResponseFreqKnife
W = 0.73796, p-value < 2.2e-16
```

In using the leveneTest function, looking at the CreativeRespFreq and Common column for the brick task, it does not violate the assumption of homogeneity of variance with a p value of .96. However the same variables for knife responses do violate the assumption of homogenous variance, with a p value of .009. Then, I made a scatterplot of the presence of common responses and creative response frequency to see if there were any outliers in the data for the brick/knife tasks. There seem to be a few outliers for the brick response, mostly in that some people with uncommon responses got very high creative response frequency scores from the interviewers. In the knife scatterplot, there's a lot more variance in their interviewer scores, so this just confirms that it violates the assumption of homogeneity of variance. Without correcting for any violations of assumptions, none of the relationships were significant so contrary to what I thought, the commonality of those responses did not affect interviewer rating of creativity.