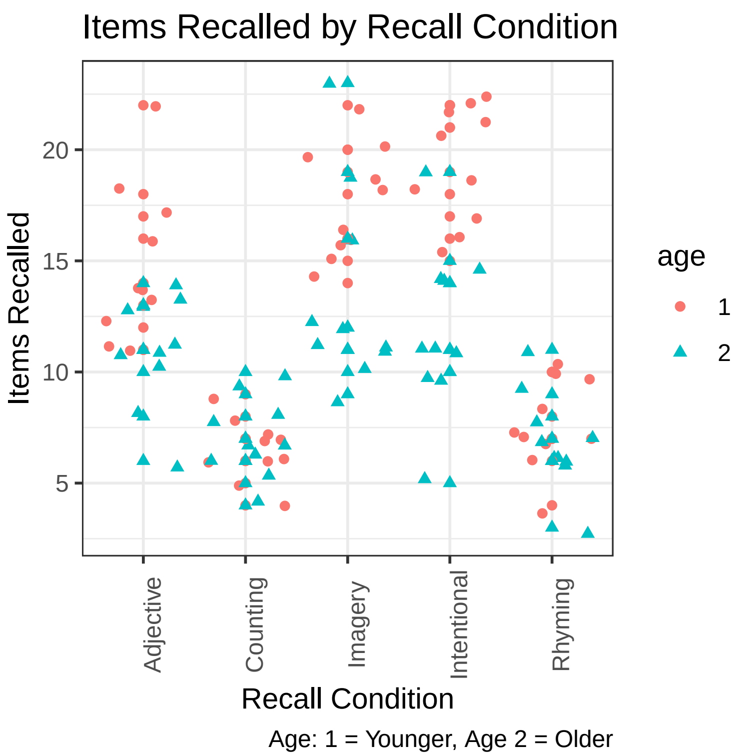
Homework Five

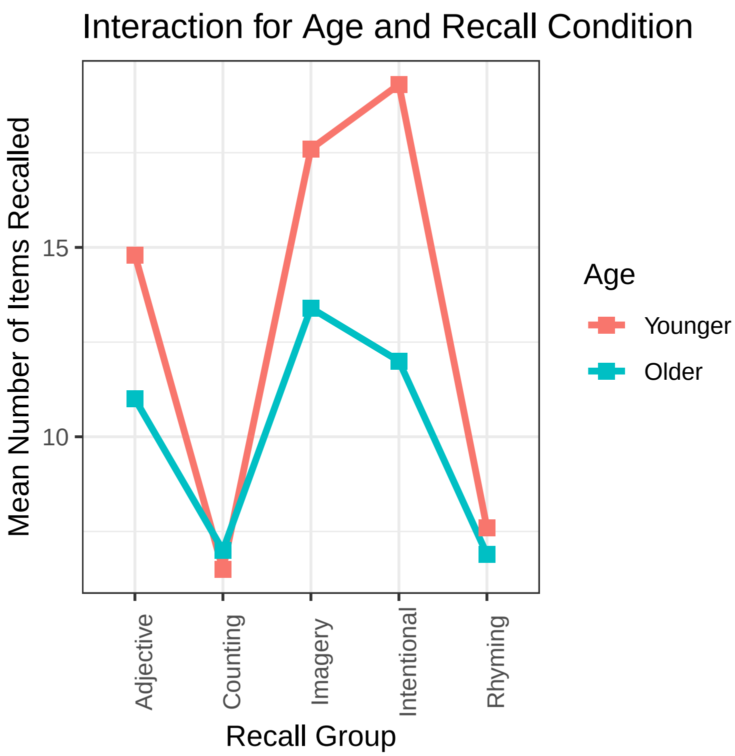
Josh Virene

Problem 1

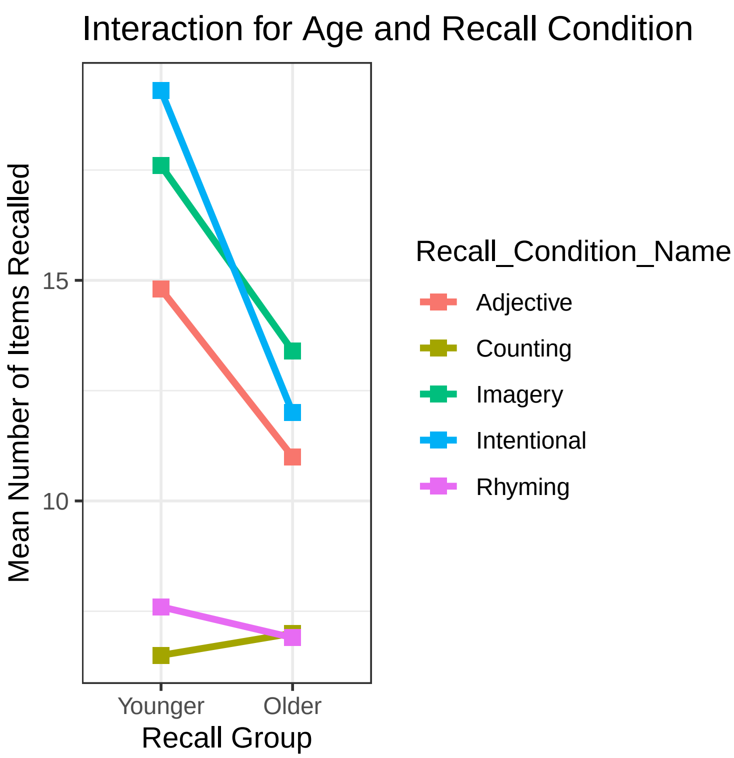
Problem 1(a)



Problem 1(b)



Problem 1(b- continued)



Between these two plots, I would report the plot with recall group as the factor on the horizontal axis, age as the factor with separate lines, and the mean of items recalled on the y axis. Looking at the R output for each plot, this plot is more interpretable between the two. These plots show the same information, but it is easier to determine whether there is interaction or not (i.e., whether the effect of recall condition depends on age) based on comparison of the slopes across age groups. It is also easier in the first plot to make comparisons across the recall groups within age groups.

Problem 1(c)

Hypothesis one: Words requiring higher levels of processing will be recalled more easily. This hypothesis suggests that the mean values for the high level of processing (adjective and imagery) will be higher than the mean values for low levels of processing (counting and rhyming). Based on the plots from 1b, this hypothesis is supported because for both the younger and older age groups, higher processing levels have higher means than those of lower processing levels. **Data:** For younger (high): $\mu_{adjective} \approx 15$, $\mu_{imagery} \approx 18$ For older (high): $\mu_{adjective} \approx 11$, $\mu_{imagery} \approx 13$ For younger (low): $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 8$ For older (low): $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 7$

Hypothesis two: Older participants will not perform as well as younger subjects at higher level processing words. Looking at the higher levels, the data support that the older participants do not perform as well as the younger participants at higher levels based on the difference in number of items recalled between age groups. **Data:** For younger: $\mu_{adjective} \approx 15$, $\mu_{imagery} \approx 18$ For older: $\mu_{adjective} \approx 11$, $\mu_{imagery} \approx 13$ These values show that the younger participants had a higher number of words recalled at the higher level than their counterparts in the older age group.

Hypothesis three: Older and younger participants will have similar levels of recall for lower level processing words. For the lower levels, the data support that the younger and older participants have similar levels of recall. The means number of words recalled are similar for the younger and older groups. **Data:** For younger: $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 8$ For older: $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 7$

Problem 1(d)

```
## Analysis of Variance Table
##
## Response: Items_Recalled
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## Recall_Condition      4 1514.94  378.73 47.1911 < 2.2e-16 ***
## Age                   1  240.25  240.25 29.9356 3.981e-07 ***
## Recall_Condition:Age  4  190.30   47.58  5.9279 0.0002793 ***
## Residuals            90  722.30    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Recall_Condition	4	1514.94	378.735000	47.191126	0.0000000
Age	1	240.25	240.250000	29.935622	0.0000004
Recall_Condition:Age	4	190.30	47.575000	5.927938	0.0002793
Residuals	90	722.30	8.025556	NA	NA

Part one, results and interpretations of significance tests:

<u>*FactorA*</u> the recall condition, is statistically significant based on its p-value of <2.2e-16. Additionally,

considering the signal to noise ratio, R has returned an F statistic of 47.1911. The interpretation of this value for the f statistic is that the variable recall condition is very important in accounting for the difference in mean number of items recalled across groups.

$\underline{FactorB}$ the age is statistically significant based on its p-value of 3.981e-07. For the signal to noise ratio, the f statistic value is 29.94. Although this is lower compared to factor A, this value indicates that age does account for some of the difference in mean number of items across groups.

$\underline{FactorC}$ the interaction for age and recall condition is statistically significant based on its p-value of 0.0002793. For the signal to noise ratio, the f statistic value is 5.93. This is the lowest of the three however, because of the p value, we can reject the null hypothesis that there is no interaction in this model (we believe the effect of age depends on the recall condition).

Part two, relation to the profile plot:

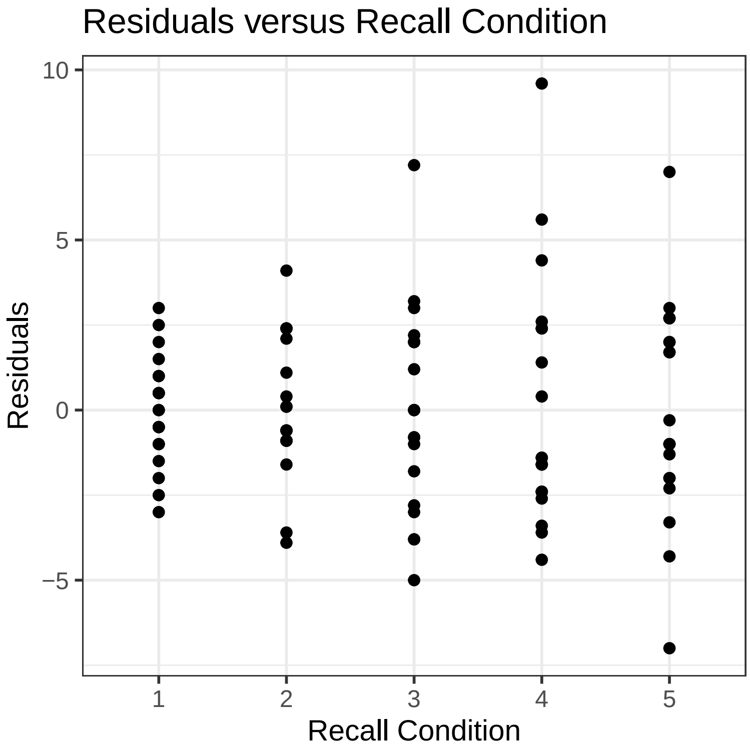
a. It is reasonable that the recall condition variable is statistically significant because for both the older and younger age groups, the mean values are different across recall conditions. Mathematically, the null hypothesis the factorial model will test for the main effect of factor A: $\tau_{adjective} \neq \tau_{counting} \neq \tau_{imagery} \neq \tau_{intentional} \neq \tau_{rhyming}$ Based on the profile plots, it makes sense that this variable is statistically significant.

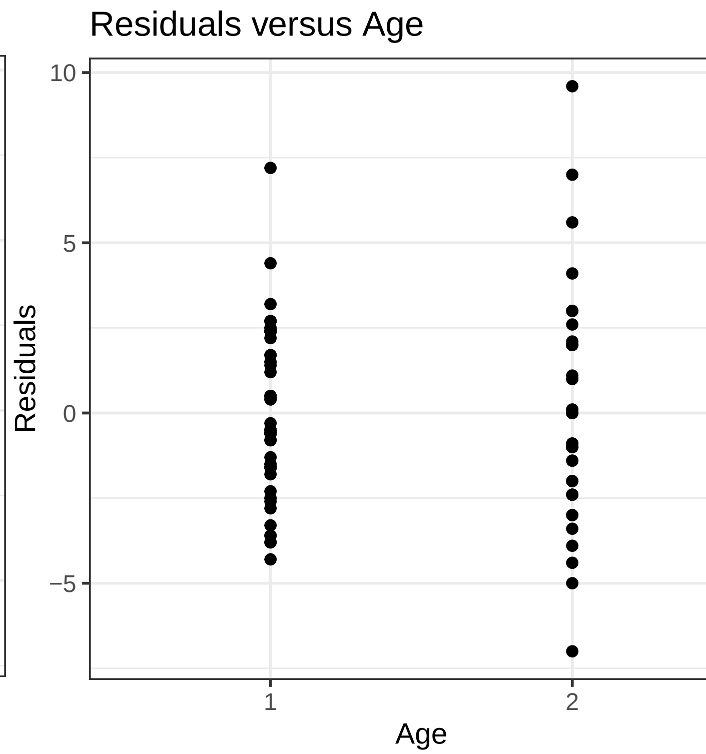
b. The age variable is statistically significant, which makes sense because across recall conditions, the means number of items recalled is different for the younger participants in the sample compared to that of the older participants. Mathematically: For younger: $\mu_{adjective} \approx 15$, $\mu_{imagery} \approx 18, \mu_{counting} \approx 7$, $\mu_{rhyming} \approx 8$ For older: $\mu_{adjective} \approx 11$, $\mu_{imagery} \approx 13$, $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 7$ Mathematically, the null hypothesis the factorial model will test for the main effect of factor B: For younger: $\beta_{younger} \neq \beta_{older}$

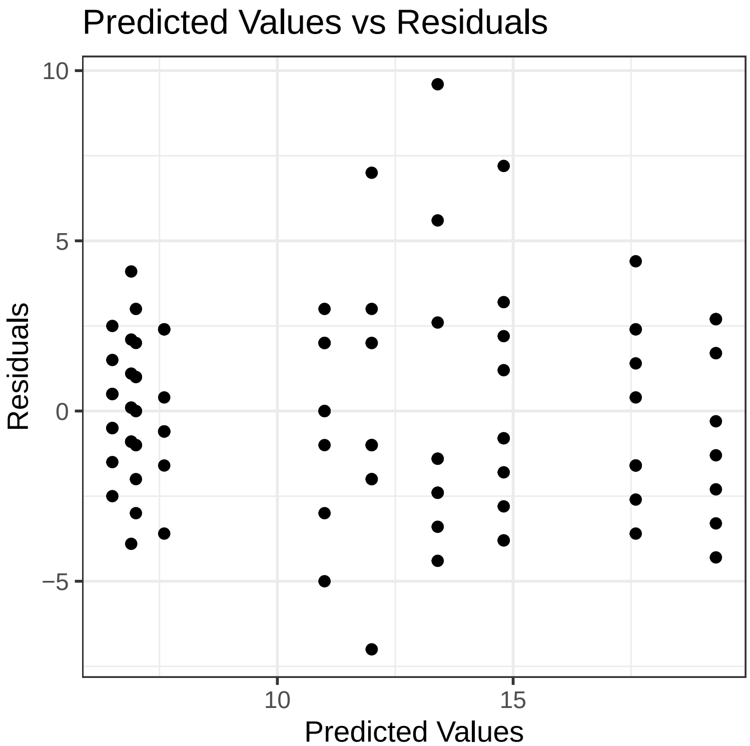
An important consideration here, is that the effect of age on items recalled is primarily exhibited by comparing the higher level of processing with the lower level. What this means is that there is a noticeable difference in number of words recalled across age groups for the high levels of processing, but not as much for the low levels, which is important and discussed below in c.

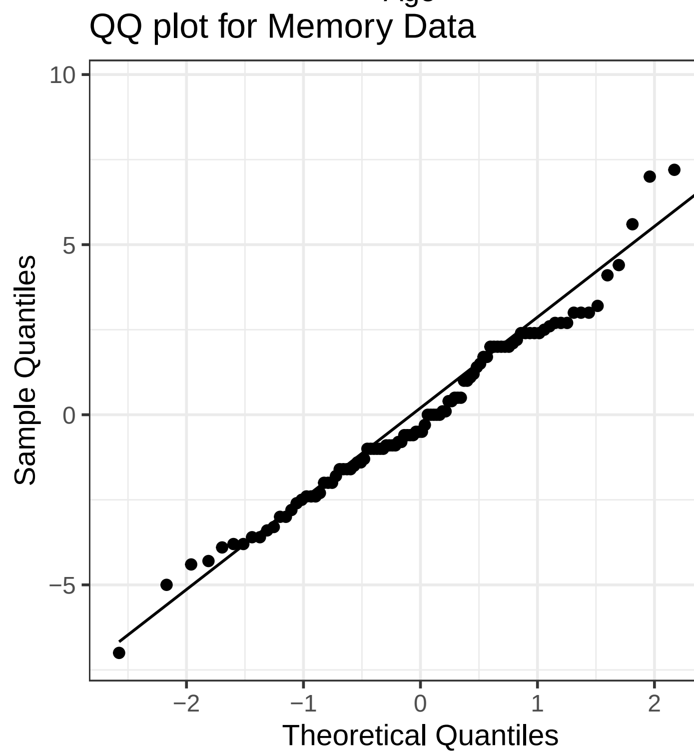
c. The interaction is statistically significant, which makes sense in the plot because the effect of recall condition depends on age. An example of this can be best described by explaining what the plot would look like without interaction between these variables, then how it does look due to the presence of interaction. In a plot without interaction, the mean number of words recalled would differ by the same amount across recall levels, and the slopes for younger and older participants in the profile plot would be parallel. Because there is interaction however, the slopes are not parallel and the effect of recall condition depends on age. Referring back to question B, it is clear that the interaction is present when comparing the higher levels of recall to lower levels due to differences in mean number of items recalled across age groups.

Problem 1(e)









a. Residuals versus recall condition: Residual plots assess the assumptions that errors / residuals are the same across different groups (equal variance). Violations of this assumption occur if variance is not equal across groups. In this plot, the groups are the recall condition groups. Based on this plot, there is a violation of this assumption, since the variance is smaller for groups one and two compared to the rest though; this violation is not extreme.

b. Residuals versus age: Residual plots assess the assumptions that errors / residuals are the same across different groups (equal variance). In this plot, this refers to the age groups, 1 = younger, 2 = older. Based on this plot, there is a violation of this assumption because the group two variance distribution is larger than the group one variance distribution.

c. Residuals versus predicted values: Residual plots assess the assumptions that errors / residuals are the same across different groups (equal variance). In this plot, this is referring to the fitted values. The assumption of equal variance across fitted values is violated.

d. QQ plot of residuals: The QQ plot assesses the assumption of normality, that the errors and residuals come from a normal distribution. Looking at this plot, the assumption is violated because there is a tail on the right end of the plot. Points will fall very close to the 1-1 line in order for this assumption to be met. The violation is not very serious in this plot.

Problem 1(f)

```
## $lsmeans
##  age lsmean    SE df lower.CL upper.CL
##  1     13.2 0.401 90    12.36     14.0
##  2     10.1 0.401 90     9.26     10.9
##
## Results are averaged over the levels of: Recall_Condition
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  1 - 2         3.1 0.567 90   5.471  <.0001
##
## Results are averaged over the levels of: Recall_Condition
```

```
## $lsmeans
##  Recall_Condition lsmean    SE df lower.CL upper.CL
##  1                  6.75 0.633 90     5.49     8.01
##  2                  7.25 0.633 90     5.99     8.51
##  3                 12.90 0.633 90    11.64    14.16
##  4                 15.50 0.633 90    14.24    16.76
##  5                 15.65 0.633 90    14.39    16.91
##
## Results are averaged over the levels of: age
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  1 - 2       -0.50 0.896 90  -0.558  0.9807
##  1 - 3       -6.15 0.896 90  -6.865  <.0001
##  1 - 4       -8.75 0.896 90  -9.767  <.0001
##  1 - 5       -8.90 0.896 90  -9.935  <.0001
##  2 - 3       -5.65 0.896 90  -6.307  <.0001
##  2 - 4       -8.25 0.896 90  -9.209  <.0001
##  2 - 5       -8.40 0.896 90  -9.377  <.0001
##  3 - 4       -2.60 0.896 90  -2.902  0.0367
##  3 - 5       -2.75 0.896 90  -3.070  0.0231
```

```
##  4 - 5       -0.15 0.896 90  -0.167  0.9998
##
## Results are averaged over the levels of: age
## P value adjustment: tukey method for comparing a family of 5 estimates
```

These comparisons do not show whether or not there is an interaction because each comparison holds the other factor constant at an average level. For example in the pairwise comparison for age, it compares averages for age while the recall condition is held constant. For the pairwise comparison for recall condition, it compares averages for recall condition while age is held constant.

Problem 1(g)

```
## $lsmeans
## age = 1:
##  Recall_Condition lsmean    SE df lower.CL upper.CL
##  1                   6.5 0.896 90     4.72     8.28
##  2                   7.6 0.896 90     5.82     9.38
##  3                  14.8 0.896 90    13.02    16.58
##  4                  17.6 0.896 90    15.82    19.38
##  5                  19.3 0.896 90    17.52    21.08
##
## age = 2:
##  Recall_Condition lsmean    SE df lower.CL upper.CL
##  1                   7.0 0.896 90     5.22     8.78
##  2                   6.9 0.896 90     5.12     8.68
##  3                  11.0 0.896 90     9.22    12.78
##  4                  13.4 0.896 90    11.62    15.18
##  5                  12.0 0.896 90    10.22    13.78
##
## Confidence level used: 0.95
##
## $contrasts
## age = 1:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2        -1.1 1.27 90  -0.868  0.9077
##  1 - 3        -8.3 1.27 90  -6.551  <.0001
##  1 - 4       -11.1 1.27 90  -8.761  <.0001
##  1 - 5       -12.8 1.27 90 -10.103  <.0001
##  2 - 3        -7.2 1.27 90  -5.683  <.0001
##  2 - 4       -10.0 1.27 90  -7.893  <.0001
##  2 - 5       -11.7 1.27 90  -9.235  <.0001
##  3 - 4        -2.8 1.27 90  -2.210  0.1854
##  3 - 5        -4.5 1.27 90  -3.552  0.0054
##  4 - 5        -1.7 1.27 90  -1.342  0.6660
##
## age = 2:
##  contrast estimate   SE df t.ratio p.value
##  1 - 2         0.1 1.27 90   0.079  1.0000
##  1 - 3        -4.0 1.27 90  -3.157  0.0180
##  1 - 4        -6.4 1.27 90  -5.052  <.0001
##  1 - 5        -5.0 1.27 90  -3.947  0.0014
##  2 - 3        -4.1 1.27 90  -3.236  0.0143
##  2 - 4        -6.5 1.27 90  -5.131  <.0001
```

```
##  2 - 5        -5.1 1.27 90  -4.025  0.0011
##  3 - 4        -2.4 1.27 90  -1.894  0.3279
##  3 - 5        -1.0 1.27 90  -0.789  0.9331
##  4 - 5         1.4 1.27 90   1.105  0.8035
##
## P value adjustment: tukey method for comparing a family of 5 estimates
```

Hypothesis one: Words requiring higher levels of processing will be recalled more easily. Based on this output, there is a large difference in means when comparing the low levels of word processing to the higher levels of word processing. This supports this null hypothesis because the number of words recalled at low levels should be low compared to that of the higher levels. An example of this is comparing group 1- counting to group 4- imagery. The difference in means for age group 1 - younger was -11.1, and for older -6.4. Furthermore, the output shows higher mean values for number of items recalled at the higher recall levels compared to the lower recall levels for both age groups in the $lsmeans section.

Hypothesis two: Older participants will not perform as well as younger subjects at higher level processing words. This null hypothesis is evaluated using the $lsmeans output, where mean values by recall condition are listed. The mean values for condition 3- adjective and for condition 4- imagery are higher for the younger participants compared to the older participants at the higher levels. Adjective: (younger = 14.8, older = 11.2) Imagery: (younger = 17.7, older = 13.4)

Hypothesis three: Older and younger participants will have similar levels of recall for lower level processing words. The data support this null hypothesis; mean values for words recalled are very similar for the younger and older participant groups for condition 1- counting (younger = 6.5, older = 7) and condition 2- rhyming (younger = 7.6, older = 6.9)

Problem 1(h)

Comparing the different recall condition groups to the *intentional* plot.

Lower levels of processing: Based on the statistical output (i.e., Tukey's HSD, emmeans), and the profile plot, telling participants to use rhyming / counting as their strategy to memorize words will yield a lower number of words recalled compared to if they were told to intentionally memorize the words. Under these two approaches, there is not a large difference in number of words called across the different age groups- **Age does not play a role**. The mean number of words memorized for the rhyming and counting groups were close for group 1 (18-30) and group 2 (55-65). For these groups there is little (if any) interaction between age and recall condition. For younger: $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 8$ For older: $\mu_{counting} \approx 7$, $\mu_{rhyming} \approx 7$

Higher levels of processing: Based on the statistical output (i.e., Tukey's HSD, emmeans), telling participants to use adjectives / imagery as their strategy for memorizing will in some cases increase the number of words recalled compared to the strategies of the lower levels of processing. This depends on which strategy is chosen as well as the age group being considered. For the adjective strategy, this did not increase number of words recalled compared to intentionally memorizing, though mean values across strategies (by age group) are closer. For the imagery strategy, the number of words recalled was higher than intentionally memorizing for the 55-65 age group, but not for the 18-30 age group. These results suggest that there is interaction; the effect of the recall condition depends on the age group in question **Age matters for higher levels of processing**. For younger: $\mu_{adjective} \approx 15$, $\mu_{imagery} \approx 18$ For older: $\mu_{adjective} \approx 11$, $\mu_{imagery} \approx 13$

Problem 2

Problem 2(a)

This plot has a main effect for location so the Location factor will be significant. This is because the mean values shown on the plot vary largely between the two locations across all treatment types. It is harder to say whether there is a main effect for treatment group because the mean values are closer together when location is constant (i.e., either one or two) so this variable may not be significant. Lastly, there is interaction in this model because the slopes are not parallel and in fact; they are very different from each other. This means interaction is significant in this model.

Problem 2(b)

In this plot, there is a main effect for location and this variable will be statistically signficant. This is because the mean value for location is different when holding treatment level constant. The variable for treatment has a main effect; location is held constant and the means across treatments are different. There is interaction in this plot because the slopes are nonparallel and the interaction variable will be significant.

Problem 2(c)

There is a main effect for Location, this is because when holding the treatment factor constant, the means are different for location one and location two across all treatment factors. The Location variable will be significant. There is a main effect for Treatment, this is because the means differ across treatment group while holding the location variable constant (i.e, for location one, the means between groups A,B, and C are different). This variable will be significant. There is not interaction in this model because the lines are parallel. It is clear that for both location one and two, treatment group B yields higher values for the dependent variable (variable of interest) compared to treatment groups A and C.

Problem 2(d)

There is a main effect in this model for Location, so this variable is significant. When holding treatment constant, the mean values are different. There is not a main effect for treatment so this variable is not significant. This is because the means are the same when holding location constant.
The lines in this plot are not parallel and in fact, they intersect which suggests that there is interaction in this model so the interaction variable will be significant.

Appendix

```
library(knitr)
#install the tidyverse library (do this once)
#install.packages("tidyverse")
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4, fig.height = 4, ti
#code for prob1a

##import dataset
library(readr)
memory <- read_csv("memory.csv")
library(ggplot2)

#convert our data to factors
memory$age <- as.factor(memory$Age)
```

```r
levels(memory$Age) <- c("Younger", "Older")

ggplot(memory) +
  theme_bw() +
  geom_jitter() +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90)) +
  aes(x = Recall_Condition_Name,
      y = Items_Recalled,
      shape = age,
      color = age) +
  labs(
    title = "Items Recalled by Recall Condition",
    caption = "Age: 1 = Younger, Age 2 = Older",
    x = "Recall Condition",
    y = "Items Recalled"
  )

#code for prob1b

## get the means for each group
memory$Age = as.factor(memory$Age)
memory$Recall_Condition = as.factor(memory$Recall_Condition)
#rename the factor variable Age to younger and older
levels(memory$Age) <- c("Younger", "Older")


memory %>%
  group_by(Recall_Condition_Name, Age) %>%
  summarize(trtmean = mean(Items_Recalled),
            sd = sd(Items_Recalled)) -> groupsummary2

## create interaction plots
#plot one
p1 <- ggplot(groupsummary2, aes(Recall_Condition_Name, trtmean)) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_line(size = 1.2, aes(group = Age, color = Age)) +
  geom_point(size = 2.6, aes(color = Age), shape = 15) +
  labs(
    title = "Interaction for Age and Recall Condition",
    x = "Recall Group",
    y = "Mean Number of Items Recalled"
  )
p1
#code for probb- continued
#plot two
p2 <- ggplot(groupsummary2, aes(Age, trtmean)) +
  theme_bw() +
  geom_line(size = 1.2, aes(group = Recall_Condition_Name, color = Recall_Condition_Name)) +
  geom_point(size = 2.6, aes(color = Recall_Condition_Name), shape = 15) +
  labs(
    title = "Interaction for Age and Recall Condition",
```

```
    x = "Recall Group",
    y = "Mean Number of Items Recalled"
  )
p2
#code for prob1c

#code for prob1d
model <- lm(Items_Recalled ~ Recall_Condition + Age + Recall_Condition*Age, data = memory)
anova <- anova(model)
(anova)
kable(anova)
#code for prob1e
model <- lm(Items_Recalled ~ Recall_Condition + age + Recall_Condition*age, data = memory)

## plot one: residuals versus recall condition
ggplot(model, aes(x=Recall_Condition,y=model$residuals)) +
  theme_bw() +
  geom_point() +
  ggtitle("Residuals versus Recall Condition") +
  xlab("Recall Condition") + ylab("Residuals")
## plot two: residuals versus age
ggplot(model, aes(x=age,y=model$residuals)) +
  theme_bw() +
  geom_point() +
  ggtitle("Residuals versus Age") +
  xlab("Age") + ylab("Residuals")
## plot three: residuals versus predicted values
ggplot(model, aes(x=model$fitted.values,y=model$residuals)) +
  theme_bw() +
  geom_point() +
  ggtitle("Predicted Values vs Residuals") +
  xlab("Predicted Values") + ylab("Residuals")
## plot four: QQplot of residuals
ggplot(model,aes(sample=model$residuals)) +
  stat_qq() +
  stat_qq_line() +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles") +
  ggtitle("QQ plot for Memory Data") +
  theme_bw()

#code for prob1f
memory$age = as.factor(memory$age)
model <- lm(Items_Recalled ~ Recall_Condition + age + Recall_Condition*age, data = memory)

library(emmeans)
lsmeans(model, pairwise ~ "age")
lsmeans(model, pairwise ~ "Recall_Condition")

#code for prob1g

model <- lm(Items_Recalled ~ Recall_Condition + as.numeric(age) + Recall_Condition*as.numeric(age), data
lsmeans(model, pairwise ~ Recall_Condition|as.numeric(age))
```

```
#code for prob1h
```