RMHI/ARMP Problem Set 2

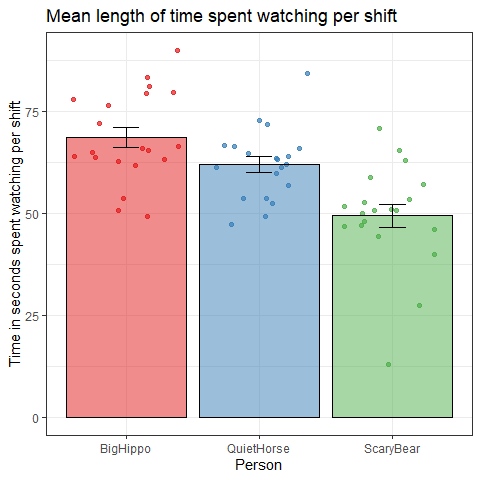
Koquiun Li Lin 1319881 [Word Count: 1630]

Please put your answers here, following the instructions in the assignment description. Put your answers and word count tallies in the locations indicated; if none is indicated that means there is no word count for that question. Remember to knit as you go, and submit the knitted version of this on Canvas.

## Q1

**Q1a**

# any code goes here  
ds\_sum <- ds %>%  
 group\_by(person) %>%  
 summarise(mean = mean(length),  
 sd = sd(length),  
 n = n(),  
 sderr = sd/sqrt(n)) %>%  
 ungroup()  
  
# bar graph with all foods (new data)  
ds\_sum %>%  
 ggplot(mapping=aes(x=person,y=mean,fill=person)) +  
 geom\_jitter(data=ds,mapping=aes(x=person,y=length,colour=person),alpha=0.7,show.legend=FALSE) +  
 geom\_col(colour="black",show.legend=FALSE,alpha=0.5) +  
 geom\_errorbar(mapping=aes(ymin = mean-sderr, ymax= mean+sderr),width=0.2) +  
 theme\_bw() +  
 scale\_fill\_brewer(palette="Set1") +  
 scale\_colour\_brewer(palette="Set1") +  
 labs(title = "Mean length of time spent watching per shift", x="Person", y="Time in seconds spent watching per shift")



**Q1b**

# any code goes here  
#test if the residuals are normally distributed  
model1 <- aov(length ~ person, data=ds)  
shapiro.test(model1$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model1$residuals  
## W = 0.96825, p-value = 0.1198

#test if the variance is homogeneous  
leveneTest(length ~ person, data=ds)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.4662 0.6297  
## 57

*Note that the order of these assumptions (i.e., what assumption you put in 1 vs 2) does not matter!*

*ASSUMPTION 1: Residuals are normally distributed. The Shapiro-Wilk test used is not significant, W=0.97, p=.120, suggesting that it is safe to assume the residuals are normally distributed, since the p-value is greater than 0.05, and therefore the assumption is not violated. [Word Count: 43]*

*ASSUMPTION 2: Variance is homogeneous. The Levene test used is not significant, F(2,57)=0.47, p=.630, suggesting that there is no significant difference in variances among the groups, since the p-value is greater than 0.05, and therefore the assumption is not violated. [Word Count: 44]*

**Q1c**

# any code goes here  
summary(model1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## person 2 3789 1894.6 15.9 3.25e-06 \*\*\*  
## Residuals 57 6792 119.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

etaSquared(model1)

## eta.sq eta.sq.part  
## person 0.3581173 0.3581173

*ANSWER: As the residuals were normally distributed and the variance was homogeneous, we ran a one-way ANOVA. The predictor variable was person, while the outcome variable was length. The statistical test was significant, F(2,57)=15.9, p<.0001, suggesting that there was a significant difference between the three people in the length of time spent watching. Moreover, the effect size (eta-squared) was 0.36, indicating that the grouping based on person explained 36% of the variance in the length of time spent watching. [Word Count: 84]*

**Q1d**

# any code goes here  
library(lsr)  
posthocPairwiseT(x=model1,p.adjust.method="holm")

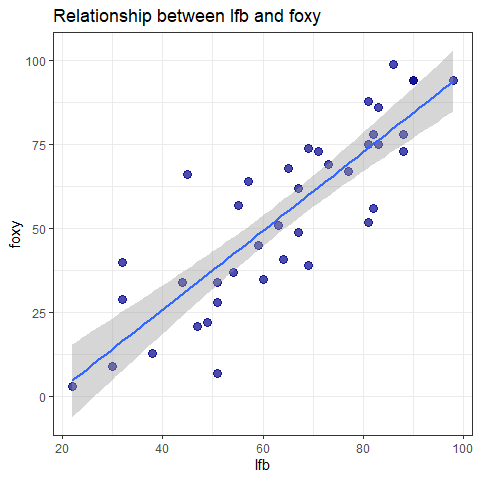
##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: length and person   
##   
## BigHippo QuietHorse  
## QuietHorse 0.0628 -   
## ScaryBear 2.3e-06 0.0011   
##   
## P value adjustment method: holm

*ANSWER: The comparison between ScaryBear and BigHippo, as well as between ScaryBear and QuietHorse, yielded significant p-values, both less than .01. These results indicated a significant difference in the mean length of time spent watching between ScaryBear and BigHippo, and between ScaryBear and QuietHorse. [Word count: 43]*

## Q2

**Q2a**

# any code goes here  
dl %>%  
 ggplot(mapping=aes(x=lfb,y=foxy)) +  
 geom\_point(colour="darkblue",alpha=0.7,size=3) +   
 theme\_bw() +  
 geom\_smooth(method="lm") +  
 labs(title = "Relationship between lfb and foxy",   
 x="lfb",   
 y="foxy")



cor.test(dl$lfb,dl$foxy)

##   
## Pearson's product-moment correlation  
##   
## data: dl$lfb and dl$foxy  
## t = 10.141, df = 38, p-value = 2.31e-12  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7399745 0.9208905  
## sample estimates:  
## cor   
## 0.8545113

*ANSWER: To measure the correlation between the answers from LFB and Foxy. We firstly observed the relationship was linear, then conducted a Pearson’s correlation test to find the correlation, which was significant, t(38)=10.14, p<.0001, with a correlation coefficient of 0.85. Therefore, the answers exhibited a strong positive relationship. [Word Count: 47]*

**Q2b**

# any code goes here  
model2 <- lm(lfb ~ foxy,data=dl)  
summary(model2)

##   
## Call:  
## lm(formula = lfb ~ foxy, data = dl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -26.5573 -6.0572 0.6602 5.8847 18.1674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.42665 3.71081 8.199 6.25e-10 \*\*\*  
## foxy 0.62319 0.06145 10.141 2.31e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.13 on 38 degrees of freedom  
## Multiple R-squared: 0.7302, Adjusted R-squared: 0.7231   
## F-statistic: 102.8 on 1 and 38 DF, p-value: 2.31e-12

*ANSWER: (i) The p-value (2.31e-12) is the same for the correlation and the linear regression; The t-value (10.141) in correlation is the same as the t-value of “foxy” in linear regression; (ii) The parameter “residual R-squared” of linear regression is related to the correlation coefficient r. Residual R-squared is a measure of effect size, calculated as (1 - SSr/SStot), where SSr is the deviation between predictions and actual values, and SStot represents the total variability in the outcome; (iii) Linear regression provides additional insights by estimating the slope and intercept parameters. The slope quantifies how the outcome variable changes with each unit change in the predictor variable, while the intercept represents the value of the outcome variable when the predictor variable is zero. [Word Count: 121]*

**Q2c**

# any code goes here  
library(lsr)  
standardCoefs(model2)

## b beta  
## foxy 0.6231913 0.8545113

*ANSWER: The standardised coefficient is 0.855.*

**Q2d**

# this code is given  
dl <- dl %>%  
 mutate(lfbZnorm = (lfb - mean(lfb))/sd(lfb),  
 foxyZnorm = (foxy - mean(foxy))/sd(foxy))  
  
# any code goes here  
cor.test(dl$lfbZnorm,dl$foxyZnorm)

##   
## Pearson's product-moment correlation  
##   
## data: dl$lfbZnorm and dl$foxyZnorm  
## t = 10.141, df = 38, p-value = 2.31e-12  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7399745 0.9208905  
## sample estimates:  
## cor   
## 0.8545113

*ANSWER: The code calculates two new variables, “lfbZnorm” and “foxyZnorm”, by standardizing the variables “lfb” and “foxy”, respectively. Standardization transforms the variables into z-scores with a mean of 0 and a standard deviation of 1. This ensures that the variables are on a common scale, making them more comparable. While the process of standardization linearly transforms the variables, it does not change the correlation coefficient (0.85). The correlation coefficient is preserved because standardization preserves the relative relationships between the variables. Therefore, the correlation coefficient between “lfbZnorm” and “foxyZnorm” will be the same as the correlation coefficient between the original “lfb” and “foxy” variables. [Word Count: 103]*

**Q2e**

# any code goes here  
model2\_mod <- lm(lfbZnorm ~ foxyZnorm,data=dl)  
summary(model2\_mod)

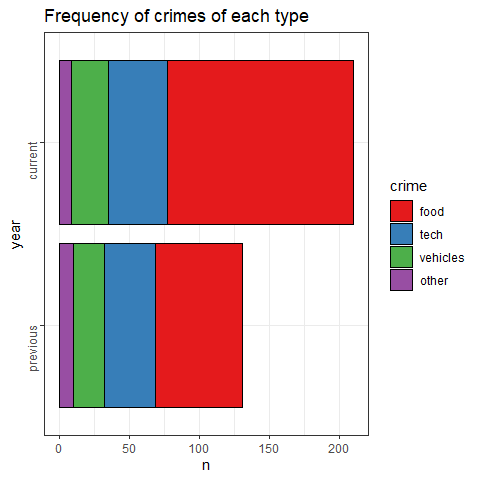
##   
## Call:  
## lm(formula = lfbZnorm ~ foxyZnorm, data = dl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.38010 -0.31477 0.03431 0.30581 0.94411   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.127e-16 8.320e-02 0.00 1   
## foxyZnorm 8.545e-01 8.426e-02 10.14 2.31e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5262 on 38 degrees of freedom  
## Multiple R-squared: 0.7302, Adjusted R-squared: 0.7231   
## F-statistic: 102.8 on 1 and 38 DF, p-value: 2.31e-12

*ANSWER: In the standardized regression, the intercept is 0 and the slope is 0.85. These values differ from those in Q2b due to standardization. Standardization shifts the mean of both outcome and predictor variables to 0, which makes the intercept significantly smaller. The slope, which measures how the outcome variable changes with a one standard deviation change in the predictor variable, is equal to the correlation coefficient, since it quantifies the linear relationship between variables in terms of their standard deviations. [Word Count: 81]*

## Q3

**Q3a**

# this code is given  
dcPlotTable <- dc %>%  
 group\_by(year,crime) %>%  
 summarise(n = n()) %>%  
 ungroup()  
  
# any code goes here  
dcPlotTable %>%  
 ggplot(mapping=aes(x=n,y=year)) +  
 geom\_col(aes(fill=crime),colour="black") +  
 theme\_bw() +  
 scale\_fill\_brewer(palette="Set1") +  
 labs(title = "Frequency of crimes of each type", x="n", y="year") +  
 theme(axis.text.y = element\_text(angle=90, vjust=0.5, hjust=1))



**Q3b**

# any code goes here  
currentTibble <- dc %>%  
 filter(year == "current")   
  
currentTable <- table(currentTibble$crime)  
currentTable

##   
## food tech vehicles other   
## 133 42 26 9

chisq.test(x=currentTable, p=rep(1/4,4))

##   
## Chi-squared test for given probabilities  
##   
## data: currentTable  
## X-squared = 174.95, df = 3, p-value < 2.2e-16

*ANSWER: We used a chi-squared goodness of fit test to examine whether the number of thefts in the food category significantly differs from the other categories in the current year. We found significant deviations, X2(3)=174.95, p<.0001. This suggests that the distribution of thefts across categories is not uniform, with food experiencing a significantly higher number of thefts compared to other categories. [Word Count: 65]*

**Q3c**

# any code goes here  
crimeTable <- table(dc$year,dc$crime)  
ctResult <- chisq.test(crimeTable)  
ctResult

##   
## Pearson's Chi-squared test  
##   
## data: crimeTable  
## X-squared = 8.7196, df = 3, p-value = 0.03326

*ANSWER: To compare the distribution of thefts between the current and previous year, we ran a Chi-squared test of independence. The results were significant, X2(3)=8.72, p=.033, suggesting that the distribution of thefts has significantly changed from the previous year to the current year. [Word Count: 47]*

**Q3d**

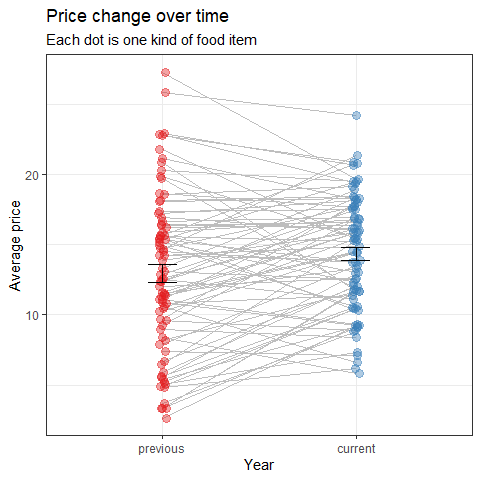
# any code goes here  
ctResult$stdres

##   
## food tech vehicles other  
## previous -2.905272 1.755136 1.139715 1.310952  
## current 2.905272 -1.755136 -1.139715 -1.310952

*ANSWER: The adjusted residuals of food is greater than the rule of thumb of 2, suggesting that food crime is significantly different between the current and the previous year. While the adjusted residuals of tech, vehicles and other are less than the rule of thumb of 2, indicating that these crime types are not significantly different between the current and the previous year. [Word Count: 62]*

## Q4

# plot the figure   
dfLong <- df %>%  
 pivot\_longer(cols=c(previous,current),names\_to="year",  
 values\_to="price")  
dfLong$year <- factor(dfLong$year,  
 levels=c("previous","current"))  
  
dfSum <- dfLong %>%  
 group\_by(year) %>%  
 summarise(mean=mean(price),  
 sd=sd(price),  
 n=n(),  
 sdErr=sd/sqrt(n)) %>%  
 ungroup()  
  
dfLong %>%  
 ggplot(mapping=aes(y=price,x=year,colour=year)) +  
 geom\_path(aes(group=names),colour="grey",  
 linewidth=0.3) +  
 geom\_jitter(alpha=0.4,size=3,width=0.02,height=0) +  
 geom\_errorbar(data=dfSum,width=0.15,  
 colour="black",  
 mapping=aes(y=mean,  
 ymin = mean-sdErr,  
 ymax = mean+sdErr)) +  
 theme\_bw() +  
 theme(legend.position="none") +  
 scale\_fill\_brewer(palette="Set1") +  
 scale\_colour\_brewer(palette="Set1") +  
 labs(title="Price change over time",  
 subtitle="Each dot is one kind of food item",  
 y="Average price", x="Year")



**Q4a**

# any code goes here  
shapiro.test(df$previous-df$current)

##   
## Shapiro-Wilk normality test  
##   
## data: df$previous - df$current  
## W = 0.97522, p-value = 0.1223

*ANSWER: A Shapiro-Wilk test assessed the assumption of normality for the difference of “previous” and “current” variables, since we are doing a paired-sample t-test. Results were not significant, W=0.98, p=.122, so normality assumptions are not violated for either year. [Word Count: 41]*

**Q4b**

# any code goes here  
t.test(x=df$previous,y=df$current,paired=TRUE)

##   
## Paired t-test  
##   
## data: df$previous and df$current  
## t = -3.0119, df = 79, p-value = 0.003486  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -2.3013185 -0.4699315  
## sample estimates:  
## mean difference   
## -1.385625

cohensD(x=df$previous,y=df$current,method="paired")

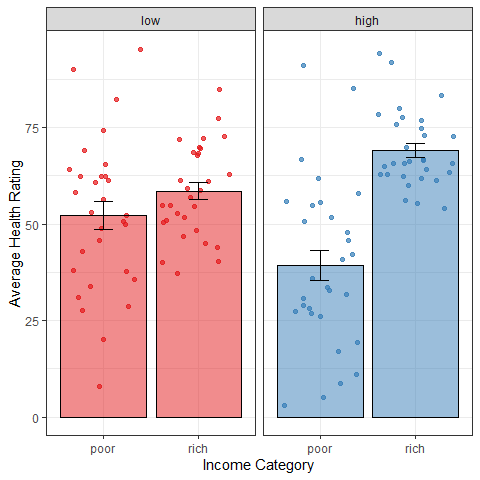
## [1] 0.3367456

*ANSWER: To analyze the change in food prices between the previous and current years, a paired t-test was conducted. This test is suitable as it compares the means of two related groups (previous and current year prices for the same food items) to ascertain any significant difference. The predictor variables are the prices of the food items in the previous and current years, and the outcome variable is the difference in prices between these two time periods. The difference between the groups was significant, t(79)=-3.01, p=.003, with a mean difference of -1.386. Moreover, the calculated effect size of 0.337 was considered as a medium effect, suggesting a moderate change in food prices between the previous and current years. [Word Count: 123]*

## Q5

**Q5a**

# any code goes here  
dh\_sum <- dh %>%  
 group\_by(inCat,exCat) %>%  
 summarise(mean = mean(health),  
 sd = sd(health),  
 n = n(),  
 sderr = sd/sqrt(n)) %>%  
 ungroup()  
  
dh\_sum %>%  
 ggplot(mapping = aes(x=inCat, y=mean, fill=exCat)) +  
 geom\_jitter(data=dh,mapping=aes(x=inCat,y=health,colour=exCat),alpha=0.7,show.legend=FALSE) +  
 geom\_col(colour="black",position="dodge",show.legend=FALSE,alpha=0.5) +  
 geom\_errorbar(mapping=aes(ymin = mean-sderr, ymax=mean+sderr),width=0.2) +  
 facet\_wrap(~exCat) +  
 theme\_bw() +  
 scale\_fill\_brewer(palette="Set1") +  
 scale\_colour\_brewer(palette="Set1") +  
 labs(x="Income Category", y="Average Health Rating", fill="Exercise Category")



**Q5b**

# any code goes here  
model3 <- aov(health ~ inCat + exCat + inCat:exCat, data=dh)  
summary(model3)

## Df Sum Sq Mean Sq F value Pr(>F)   
## inCat 1 9828 9828 35.473 2.83e-08 \*\*\*  
## exCat 1 43 43 0.156 0.693667   
## inCat:exCat 1 4177 4177 15.077 0.000172 \*\*\*  
## Residuals 116 32140 277   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*ANSWER: A two-way ANOVA test was utilized to assess the effect of two predictor variables, income category “inCat” (with poor and rich levels) and exercise category “exCat” (with low and high levels), on an outcome variable, health rating. The results showed a significant main effect on inCat, F(1,116)=35.47, p<.0001, indicating that inCat was a significantly predicted health rating. Additionally, a significant interaction effect between inCat and exCat was observed, F(1,116)=15.08, p<.001, suggesting that the effect of inCat on health rating depended on the exCat. However, there was no significant effect on exCat alone, F(1,116)=0.16, p=.694, suggesting that exCat was not a significant predictor of health rating. These findings suggested that income category and its interaction with exercise category were significant factors predicting health ratings, while exercise category alone was not a significant predictor. [Word Count: 147]*

## Q6

**Q6a**

# any code goes here  
model4 <- lm(health ~ income + exercise, data=dh)  
summary(model4)

##   
## Call:  
## lm(formula = health ~ income + exercise, data = dh)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.474 -11.639 1.287 11.410 43.464   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 24.593939 8.592920 2.862 0.00499 \*\*   
## income 0.379854 0.084126 4.515 1.52e-05 \*\*\*  
## exercise -0.007382 0.007034 -1.049 0.29613   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.27 on 117 degrees of freedom  
## Multiple R-squared: 0.154, Adjusted R-squared: 0.1396   
## F-statistic: 10.65 on 2 and 117 DF, p-value: 5.627e-05

*ANSWER: Multiple linear regression was used to evaluate a linear model with the outcome variable being the health rating, and with two predictors: the income of the person and the average amount of exercise each person got (reflected in the average calories burnt through exercise per day). The overall model was statistically significant, F(2,117)=10.65, p<.0001, with an effect size of 0.154, accounting for 15.4% of the variance in health. Income was a significant predictor: t(117)=4.52, p<.0001, with a slope of 0.38, suggesting that each additional unit of income resulted in an additional 0.38 units of health rating. However, exercise was not significant: t(117)= -1.05, p=.296, with a slope of -0.007, suggesting that an increase in calories burnt through exercise of 1kCal resulted in a decrease of 0.007 units of health rating. [Word Count: 148]*

**Q6b**

# any code goes here  
model5 <- lm(health ~ income + exercise + income:exercise, data=dh)  
summary(model5)

##   
## Call:  
## lm(formula = health ~ income + exercise + income:exercise, data = dh)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.147 -8.374 0.236 9.073 47.437   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.215e+02 1.538e+01 7.901 1.75e-12 \*\*\*  
## income -7.240e-01 1.701e-01 -4.257 4.23e-05 \*\*\*  
## exercise -1.952e-01 2.699e-02 -7.232 5.49e-11 \*\*\*  
## income:exercise 2.131e-03 2.989e-04 7.131 9.18e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.3 on 116 degrees of freedom  
## Multiple R-squared: 0.4118, Adjusted R-squared: 0.3966   
## F-statistic: 27.08 on 3 and 116 DF, p-value: 2.399e-13

*ANSWER: (i) The intersection of both income and exercise was significant: t(116)=7.13, p<.0001, with a slope of 0.002, suggesting that there was an increase in health rating of 0.002 if both income and exercise increase by 1 OR both decrease by 1; (ii) The effect size of this model including an intersection term is 0.412, suggesting that both income and exercise together account for 41.2% of the variance in health. This means that the model including the interaction term fits the data better, as it explains a larger proportion of the variance in health; (iii) The unstandardized coefficient for income has decreased from 0.38 to -0.724, and that for exercise has decreased from -0.007 to -0.195. The reason for these changes is the inclusion of the interaction term with a small coefficient, suggesting that income and exercise are dependent on each other. [Word Count: 150]*

**Q6c**

*ANSWER: The analysis in Q5b and Q6b yielded different qualitative conclusions about the factors significantly associated with health rating. In terms of the output, the ANOVA in Q5b identified income category and its interaction with exercise category as significant predictors. However, the linear regression in Q6b found income, exercise, and the interaction of income and exercise, all these three factors significant. This discrepancy can be attributed to the variable types used in different statistical tests: Q5b uses categorical, while Q6b uses numerical variables. These differences in variable types lead to variations in analytical approaches, hence resulting in different conclusions. [Word Count: 98]*

## Q7

**Q7a**

*ANSWER: (i) Option B is not possible. (ii) Option A is not possible. (iii) Option A is not possible.*

**Q7b**

*ANSWER: (i) In a chi-square test, the test statistic is calculated by summing up the squared differences between the observed and expected values, which is always non-negative. Option B shows X2=-8.69 which is negative, and p=.013. Therefore, a negative value for the test statistic is not feasible. Moreover, larger values of the chi-square test statistic tend to correspond to smaller p-values, so option A is possible; (ii) In a t-test, the t-statistic measures the size of the difference relative to the variation in the data. The degrees of freedom, typically an integer, reflect the amount of independent information available for estimating variance. In Option A, t(161.2)=5.06 shows a fractional degrees of freedom, which is unusual as it’s typically an integer; (iii) The F-test checks for significant differences among the means of three or more independent groups. A larger F-statistic typically indicates a greater difference between groups, which would correspond to a smaller p-value. In option A, F(2,132)=0.85, p<.0001. We can see a small F-statistic combined with an extremely low p-value. Hence, option A is implausible. [Word Count: 183]*

## Q8

*ANSWER: My favorite character in Bunnyland story is Foxy because Foxy is calm and smart.*