Inference_Exam1_Warren_Geither

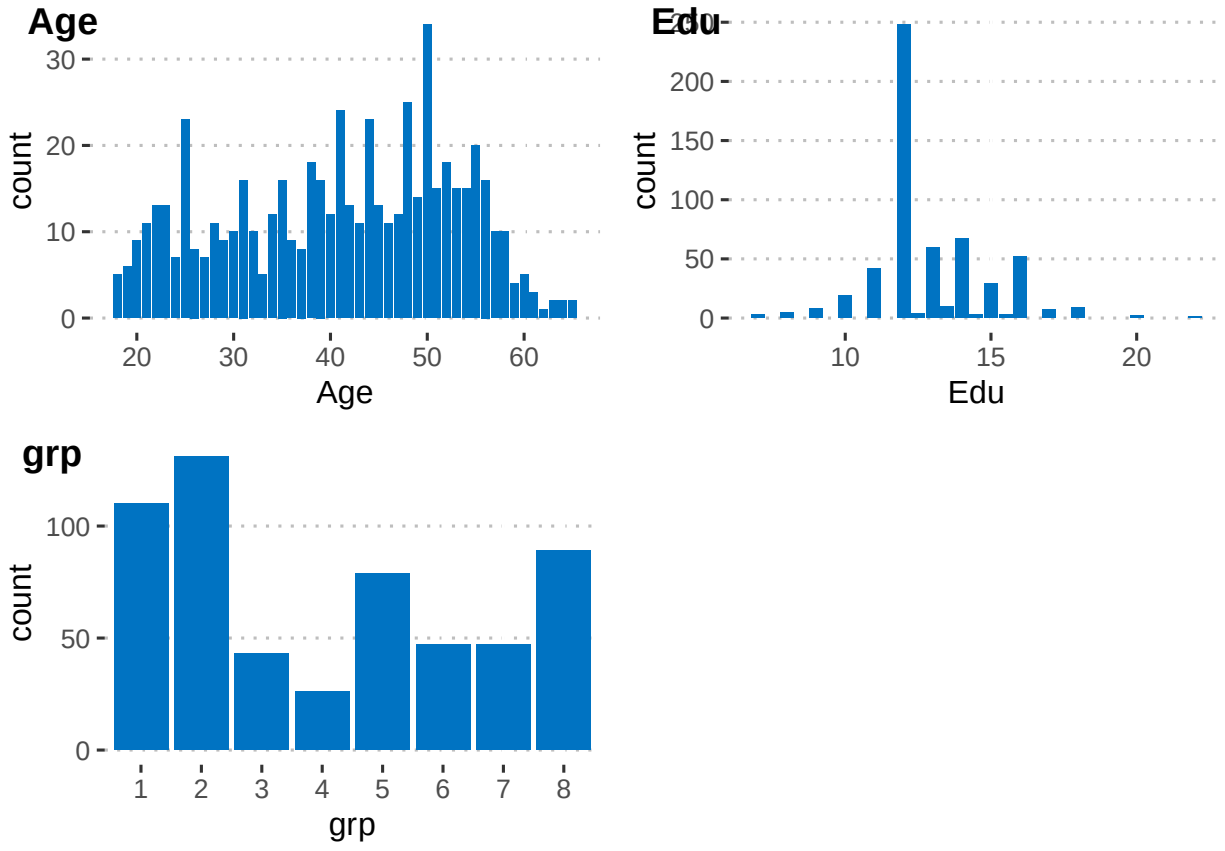
Warren Geither

10/1/2020

Taking a look at our 5 number summary. We can see see the average age for our study is around 40, average number of years in education is 12, and average delay discount is negative 4. Also notice the ranges for Age (18 - 65), Edu (7 - 22), and lnk (-12 - 4).

Age	Edu	lnk
Min. :18.00	Min. : 7.00	Min. :-12.793
1st Qu.:31.00	1st Qu.:12.00	1st Qu.: -6.343
Median :42.00	Median :12.00	Median : -4.674
Mean :40.98	Mean :12.87	Mean : -4.571
3rd Qu.:50.00	3rd Qu.:14.00	3rd Qu.: -2.935
Max. :65.00	Max. :22.00	Max. : 4.143

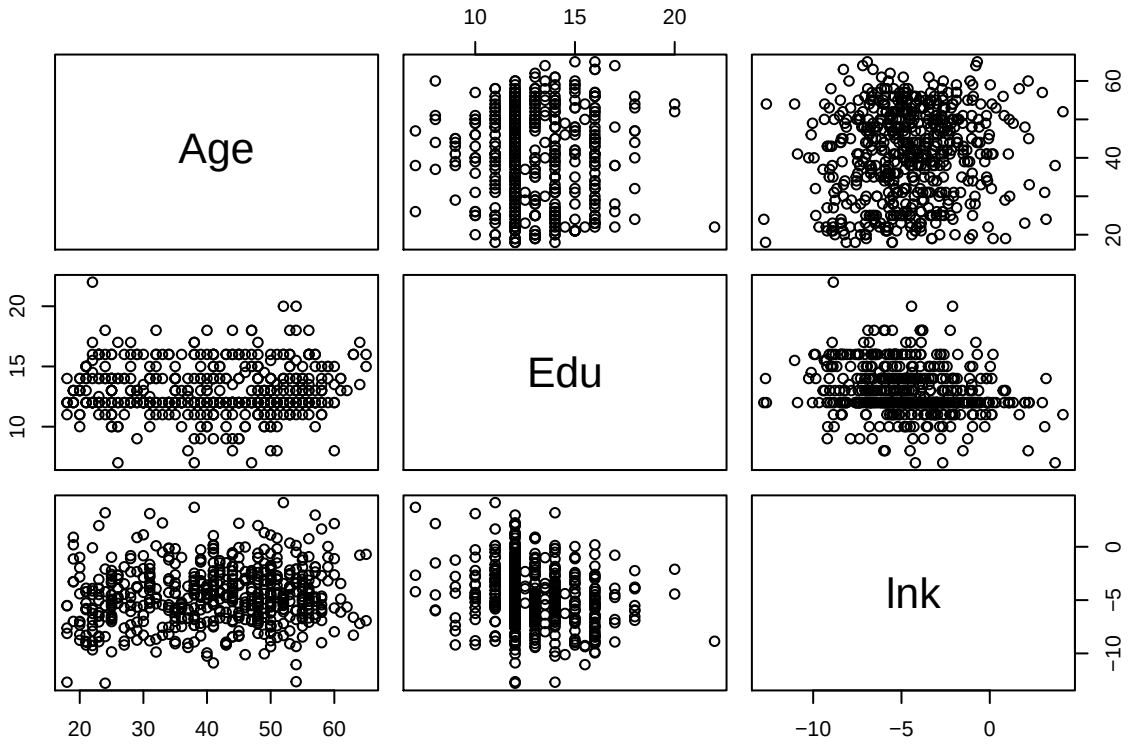
Next we'll look at the distribution of our variables.



We can see age is fairly constant with maybe a slight bias towards >=40. We can clearly see the mean and mode of Education at 12 years. And notice we have the most subjects in gropus 1 & 2 (No dependence & Cigarette only dependence).

Now, if we will be using these data in any kind of predictive/classification model , we need to make sure that there is no "multicolinearity." Meaning that our predictive variables have no relationship with each other. We need them all to be independent so that we can isolate the effect each one has on the response variable.

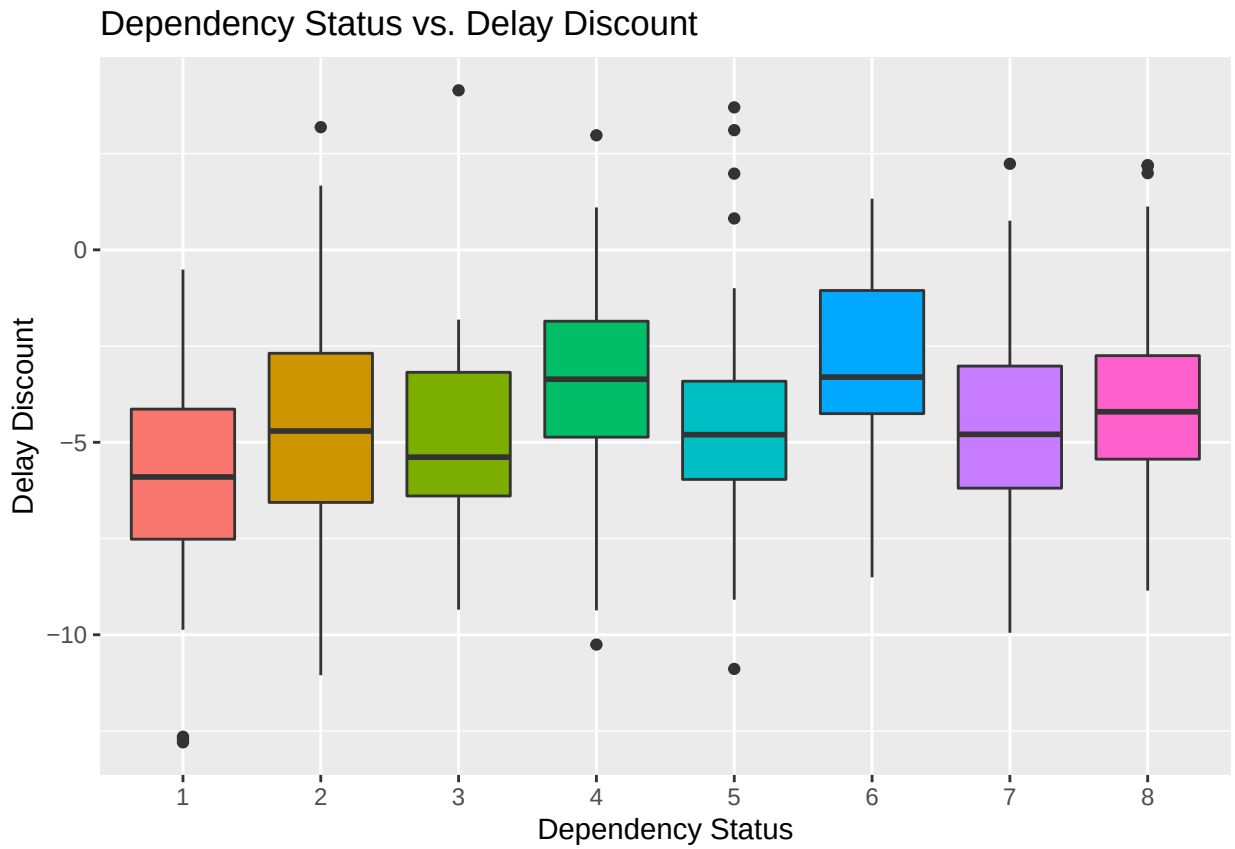
All these plots show us is that there does not appear to be any multicolinearity going on.



Next we'll look at how each of our variables relate to the response, dependency status.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

grp	min	q1	mean	median	q3	max
1	-12.793060	-7.516882	-5.913473	-5.900374	-4.137550	-0.5138452
2	-11.050832	-6.561614	-4.553586	-4.707052	-2.687023	3.1868640
3	-9.348454	-6.394950	-4.949333	-5.386056	-3.181647	4.1431347
4	-10.254464	-4.866497	-3.428741	-3.361020	-1.854393	2.9754503
5	-10.886443	-5.963880	-4.516768	-4.802402	-3.413272	3.7006248
6	-8.507243	-4.255556	-2.992683	-3.304978	-1.056793	1.3282426
7	-9.948485	-6.191632	-4.592665	-4.792705	-3.020063	2.2353763
8	-8.852666	-5.437579	-3.956524	-4.206394	-2.748872	2.2005524

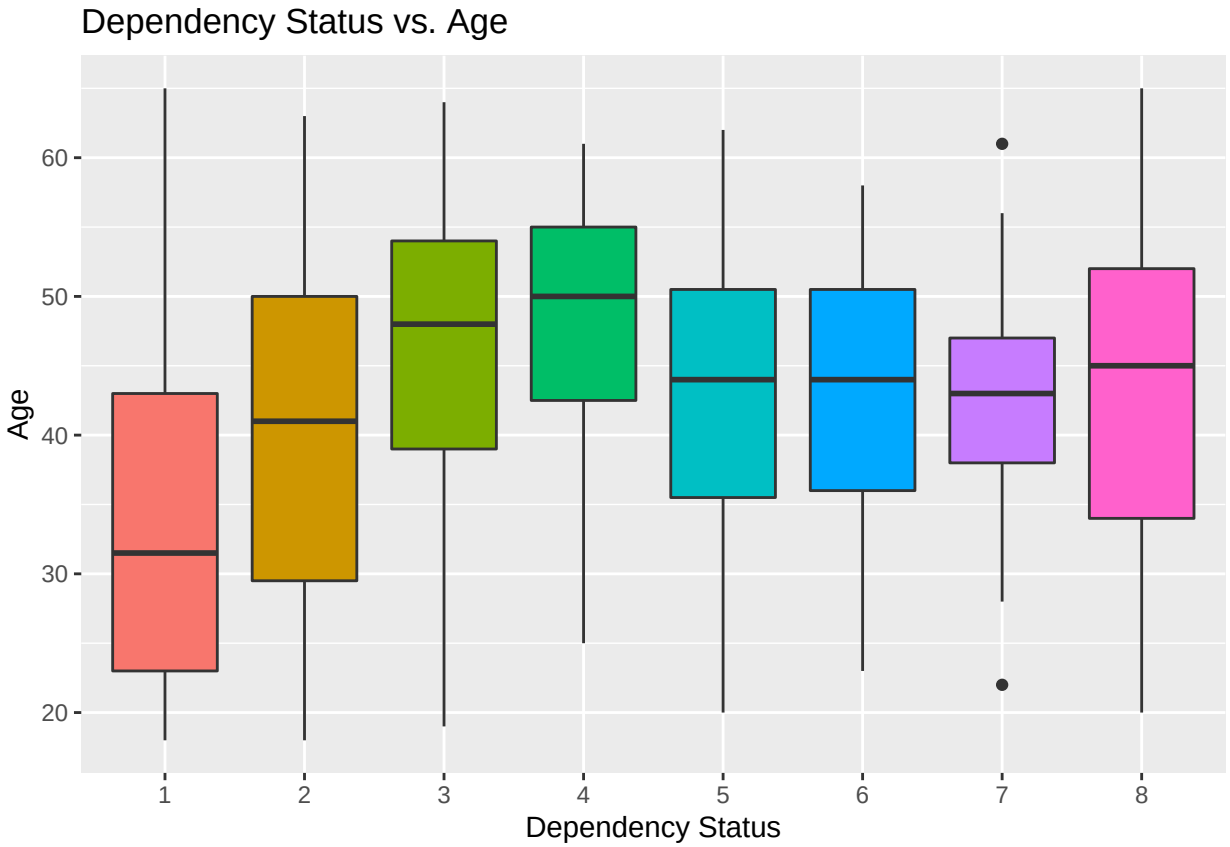


Average delay discounts hover above or below -5. Deviation from this can be seen in group 1 which has a lower median delay discount. You'll notice higher delay discounts in groups that are dependent upon stimulants (cocaine & cigarettes). And lower delay discount in group 3 when only a depressant (alcohol) is depended on.

Switching over to look at Age as it relates to Dependency Status.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

grp	min	q1	mean	median	q3	max
1	18	23.0	33.89091	31.5	43.0	65
2	18	29.5	40.42748	41.0	50.0	63
3	19	39.0	45.46512	48.0	54.0	64
4	25	42.5	46.92308	50.0	55.0	61
5	20	35.5	42.68354	44.0	50.5	62
6	23	36.0	42.97872	44.0	50.5	58
7	22	38.0	42.53191	43.0	47.0	61
8	20	34.0	43.24719	45.0	52.0	65

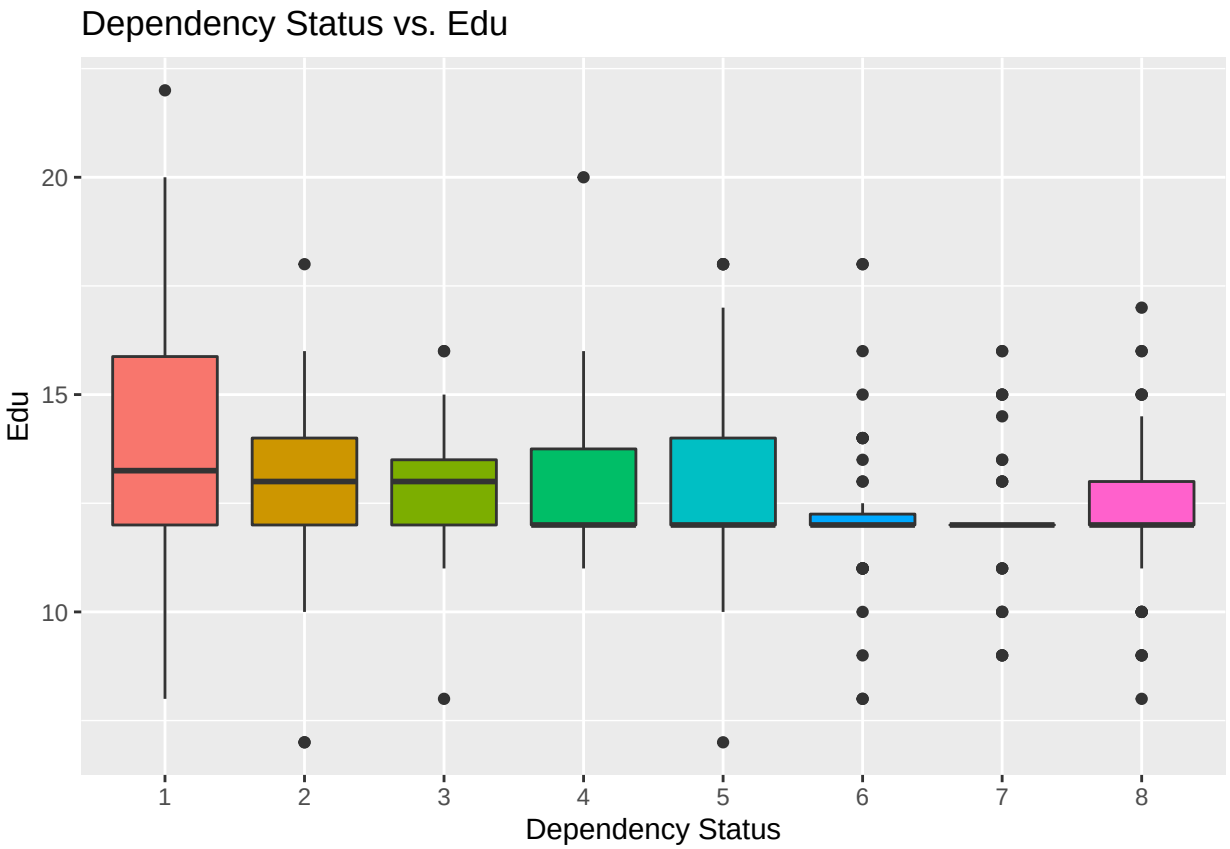


It seems subjects younger than 42 are more likely to be in the first 2 groups, while older subjects are in the latter.

Looking at education...

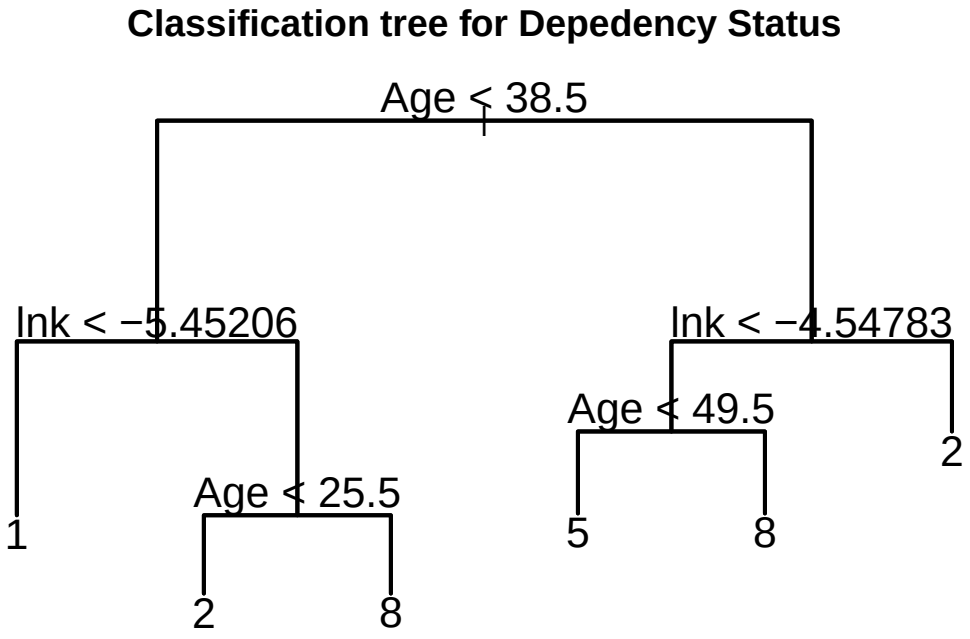
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

grp	min	q1	mean	median	q3	max
1	8	12	13.61818	13.25	15.875	22
2	7	12	13.07252	13.00	14.000	18
3	8	12	12.81395	13.00	13.500	16
4	11	12	13.09615	12.00	13.750	20
5	7	12	12.96835	12.00	14.000	18
6	8	12	12.25532	12.00	12.250	18
7	9	12	12.15957	12.00	12.000	16
8	8	12	12.24719	12.00	13.000	17



Again we see groups hovering around the average of 12 years. Also we'll see that in the first 3 groups the median education years looks slightly higher than the other 5 groups. There also seems to more spread in groups 6, 7, 8, all groups where cocaine is coupled with another substance.

b.)




```
##
## Classification tree:
## tree(formula = grp ~ ., data = data_df)
## Variables actually used in tree construction:
## [1] "Age" "lnk"
## Number of terminal nodes:  6
## Residual mean deviance:  3.631 = 2055 / 566
## Misclassification error rate: 0.6941 = 397 / 572
```

(i) Age and lnk because their branches are the biggest.

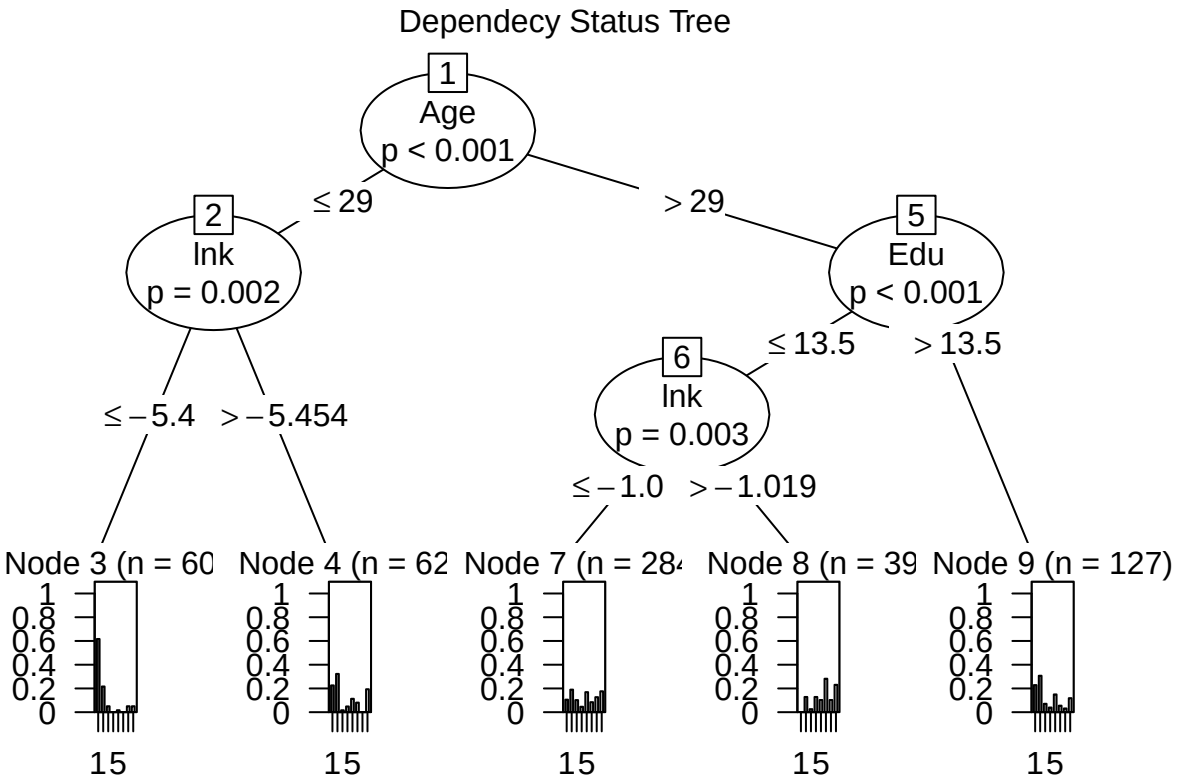
(ii) 0.6941

(iii) No, they did not fit into any of the splits defined by the tree

(iv) Broadly speaking, I think it agrees. In my EDA I said age as a split for groups 1 & 2. Also it looked like there were differences in lnk.

(v) It has a pretty high misclassification rate, so I did cross validation to see if a different size tree would help, but it already has the optimal number of branches.

c.)



(ii) There are differences. The party package uses statistical test to fit the best tree while tree just uses binary recursive partitioning. - references:

- https://cran.r-project.org/web/packages/partykit/partykit.pdf
- https://cran.r-project.org/web/packages/tree/tree.pdf

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(ggpubr)
library(tidyverse)
library(tree)
library(party)
library(partykit)
# load in file from working directory
data_df <- read.csv2("data for problem 2.csv", sep=",")

# dim of data
dim(data_df)

# check it out
```

```r
head(data_df)

# change Edu column to numeric since its number of years
data_df$Edu <- as.numeric(data_df$Edu)

# change lnk column to numeric since its log number
data_df$lnk <- as.numeric(data_df$lnk)

# change frp column to character since its non-numeric
data_df$grp <- as.factor(data_df$grp)

levels(data_df$grp) <- c("No dependency"
                         ,"Cigarette dependence only"
                         ,"Alcohol dependence only"
                         ,"Cocaine dependence only"
                         ,"Cigarette and alcohol dependence"
                         ,"Cigarette and cocaine dependence"
                         ,"Alcohol and cocaine dependence"
                         ,"Cigarette, alcohol, and cocaine dependence")

# label the levels so they plot nice
levels(data_df$grp) <- c("1"
                         ,"2"
                         ,"3"
                         ,"4"
                         ,"5"
                         ,"6"
                         ,"7"
                         ,"8")
# print summary
knitr::kable(summary(data_df[1:3]))
# frequency of Age
plot1 <- ggplot(data=data_df, aes(Age)) +
          geom_bar(fill="#0073C2FF") +
          theme_pubclean()

# frequency of Edu
plot2 <- ggplot(data=data_df, aes(Edu)) +
          geom_bar(fill="#0073C2FF") +
          theme_pubclean()

# frequency of grp
plot3 <- ggplot(data=data_df, aes(grp)) +
          geom_bar(fill="#0073C2FF") +
          theme_pubclean()

# plot all on same page
ggarrange(plot1, plot2, plot3,
          labels = c("Age", "Edu", "grp"),
          ncol = 2, nrow = 2)
# scatter plot matrix of age, edu, and lnk
pairs(data_df[1:3])
# summary table of lnk by group
```

```r
lnk_summary_df <- data_df %>%
                    group_by(grp) %>%
                    summarize(min = min(lnk)
                              , q1 = quantile(lnk, 0.25)
                              , mean = mean(lnk)
                              , median = median(lnk)
                              , q3 = quantile(lnk, 0.75)
                              , max = max(lnk))

# print summary table
knitr::kable(lnk_summary_df)

# boxplot of grp vs lnk
ggplot(data_df, aes(x=grp, y=lnk, fill=grp)) +
    geom_boxplot() +
    ggtitle("Dependency Status vs. Delay Discount") +
    labs(x="Dependency Status", y="Delay Discount") +
    theme(legend.position = "none")
# summary of age by group
age_summary_df <- data_df %>%
                    group_by(grp) %>%
                    summarize(min = min(Age)
                              , q1 = quantile(Age, 0.25)
                              , mean = mean(Age)
                              , median = median(Age)
                              , q3 = quantile(Age, 0.75)
                              , max = max(Age))

# print table
knitr::kable(age_summary_df)

# boxplot of grp vs Age
ggplot(data_df, aes(x=grp, y=Age, fill=grp)) +
    geom_boxplot() +
    ggtitle("Dependency Status vs. Age") +
    labs(x="Dependency Status", y="Age") +
    theme(legend.position = "none")
# summary of edu by group
edu_summary_df <- data_df %>%
                    group_by(grp) %>%
                    summarize(min = min(Edu)
                              , q1 = quantile(Edu, 0.25)
                              , mean = mean(Edu)
                              , median = median(Edu)
                              , q3 = quantile(Edu, 0.75)
                              , max = max(Edu))

# print table
knitr::kable(edu_summary_df)

# boxplot of grp vs Edu
ggplot(data_df, aes(x=grp, y=Edu, fill=grp)) +
    geom_boxplot() +
```

```
    ggtitle("Dependency Status vs. Edu") +
    labs(x="Dependency Status", y="Edu") +
    theme(legend.position = "none")
# fit tree
data_tree <- tree(grp~., data = data_df)

# plotting tree
plot(data_tree,lwd=2)
text(data_tree,cex=1.3)
title(main='Classification tree for Depedency Status')

# get the number of misclassifications
summary(data_tree)
# run cross validation on the tree
test <- cv.tree(data_tree, FUN = prune.misclass)

# plot new optimal tree
prune_data_tree = prune.misclass(data_tree, best = 6)
#plot(prune_data_tree)
#text(data_tree,cex=1.3)
# create new tree
party_tree = ctree(grp~Age + Edu + lnk, data = data_df)
plot(party_tree, main="Dependecy Status Tree")
```