Charlie regression analysis

Carolyn Jane Anderson

2023-03-28

This notebook contains code for generating figures using the StudentEval dataset.

```
data.raw = read.csv('../raw_data/interactions.csv',header=TRUE,stringsAsFactors=FALSE)

data <- data.raw %>% mutate(success = ifelse(tests_passed==total_tests,1,0),
                            firstAttempt = ifelse(first_attempt=="True",1,0),
                            lastAttempt = ifelse(last_attempt=="True",1,0),
                            group = ifelse(first_attempt=="True"&success==1,"SuccessFirst",ifelse(first_
                                           "UnsuccessFirst",
                                           ifelse(last_attempt=="True"&success==1,
                                                  "SuccessLast",
                                                  ifelse(last_attempt=="True","UnsuccessLast","Middle"))
```

```
exclude = read.csv('exclude.csv',header=TRUE,stringsAsFactors=FALSE) %>% mutate(joined = paste(problem,s
data.sub <- data %>% mutate(join_ed = paste(problem,str_trim(submitted_text)))
remove <- subset(data.sub,join_ed %in% exclude$joined)
cleaned <- subset(data.sub,!(join_ed %in% exclude$joined))
```

```
pass.raw = read.csv('../computed_data/allprompts_starcoderbase_pass1.csv',header=TRUE)
pass <- subset(pass.raw,select=c("prompt","pass1"))
data.all <- merge(cleaned,pass,by="prompt")
```

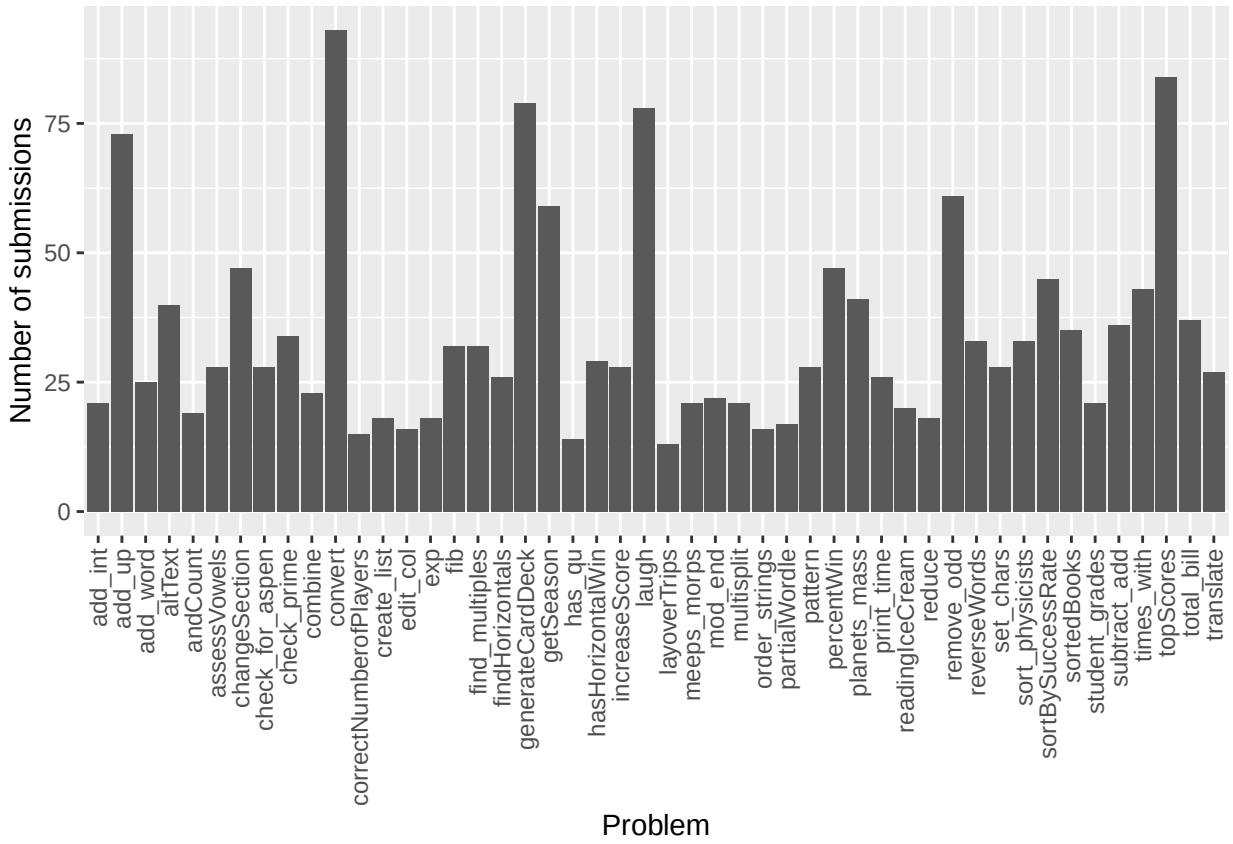
```
group_order <- c("Middle","UnsuccessLast","SuccessLast","UnsuccessFirst","SuccessFirst")

data.all$group <- factor(data.all$group,levels = group_order)

first_labs <- c(`False` = "First",`True` = "Middle/Last")
```

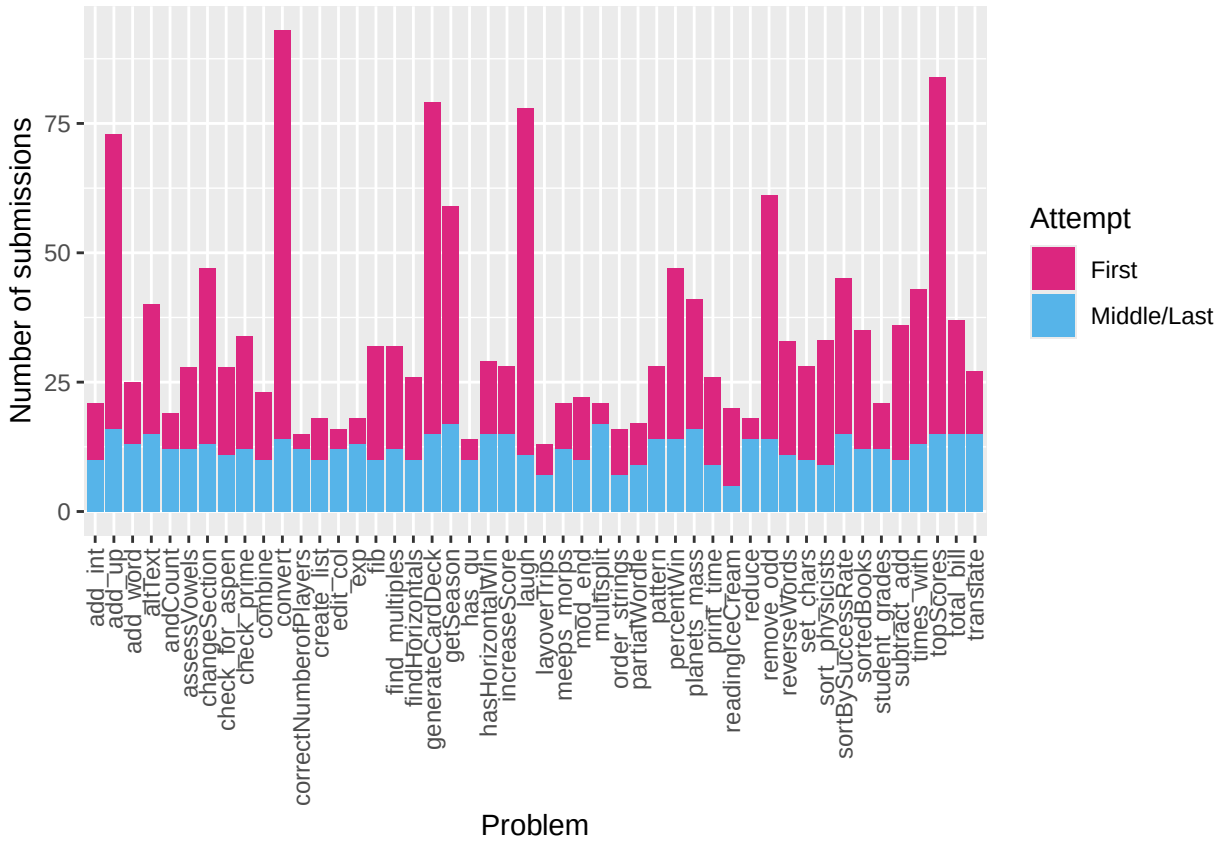
```
ggplot(data=data.all,aes(x=problem)) + geom_histogram(stat="count") + theme(axis.text.x = element_text(a
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```




```
ggplot(data=data.all,aes(x=problem,fill=first_attempt)) + geom_histogram(stat="count") + theme(axis.tex
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```



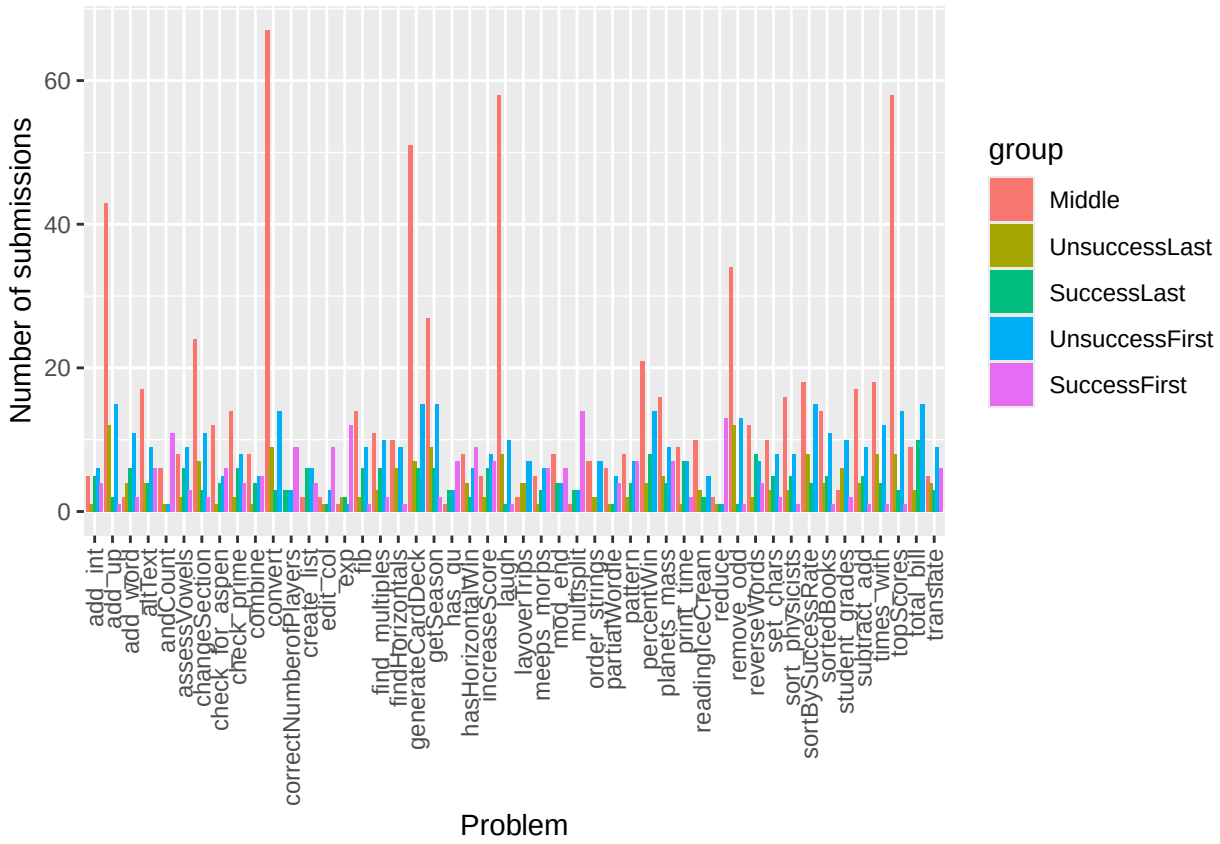

```
ggsave("n_subs.pdf")
```

```
## Saving 6.5 x 4.5 in image
```

```
counts <- data %>% group_by(problem) %>% summarize(countP = n())
succCounts <- subset(data,group=="SuccessFirst"|group=="SuccessLast") %>% group_by(problem) %>% summari
names(counts)[names(counts) == 'problem'] <- 'p'
```

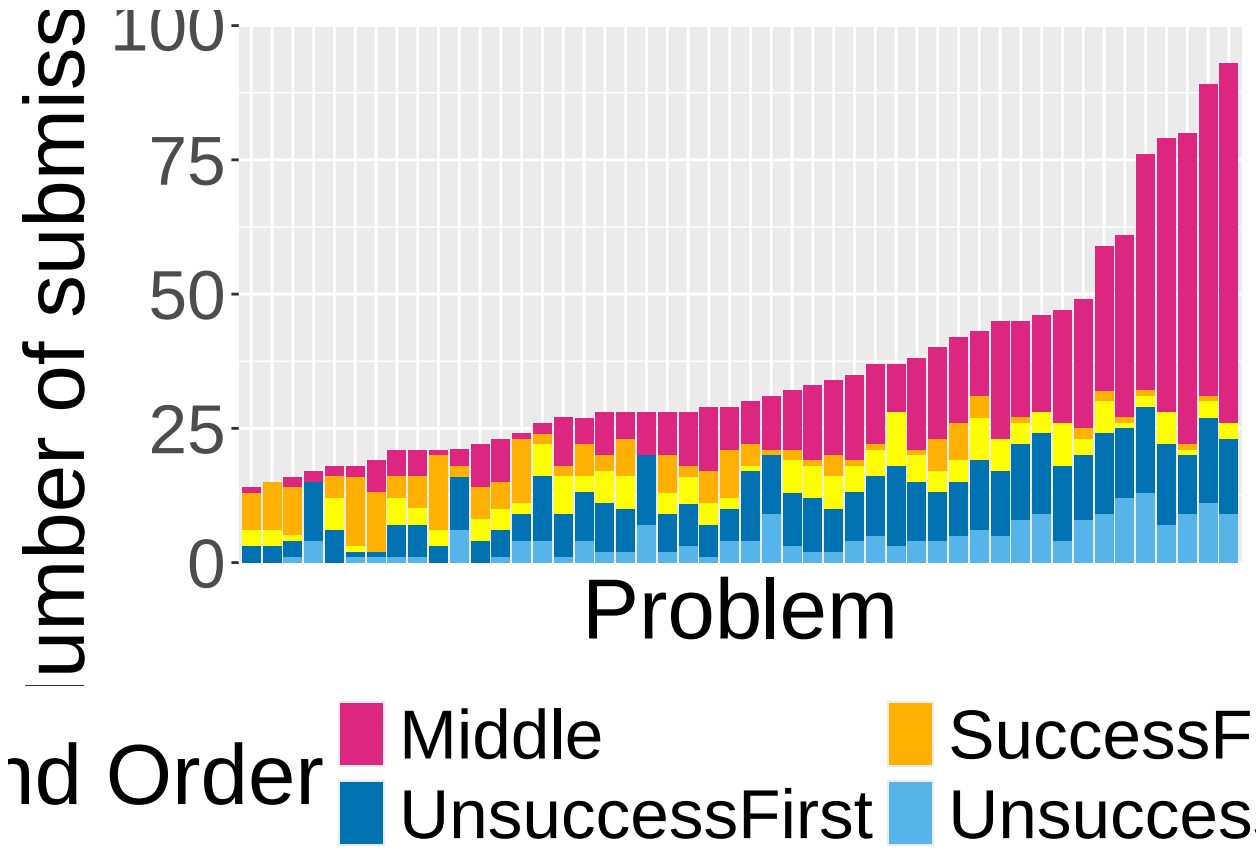
```
ggplot(data=data.all,aes(x=problem,fill=group)) + geom_histogram(stat="count",position="dodge") + theme
```

```
## Warning in geom_histogram(stat = "count", position = "dodge"): Ignoring unknown
## parameters: `binwidth`, `bins`, and `pad`
```




```
data <- data %>% rowwise() %>% mutate(countSub = subset(counts,p==problem)$countP)
data$problem <- reorder(data$problem,data$countSub)
ggplot(data=data,aes(x=problem,fill=group)) + geom_histogram(stat="count") + xlab("Problem") + ylab("Nu
```

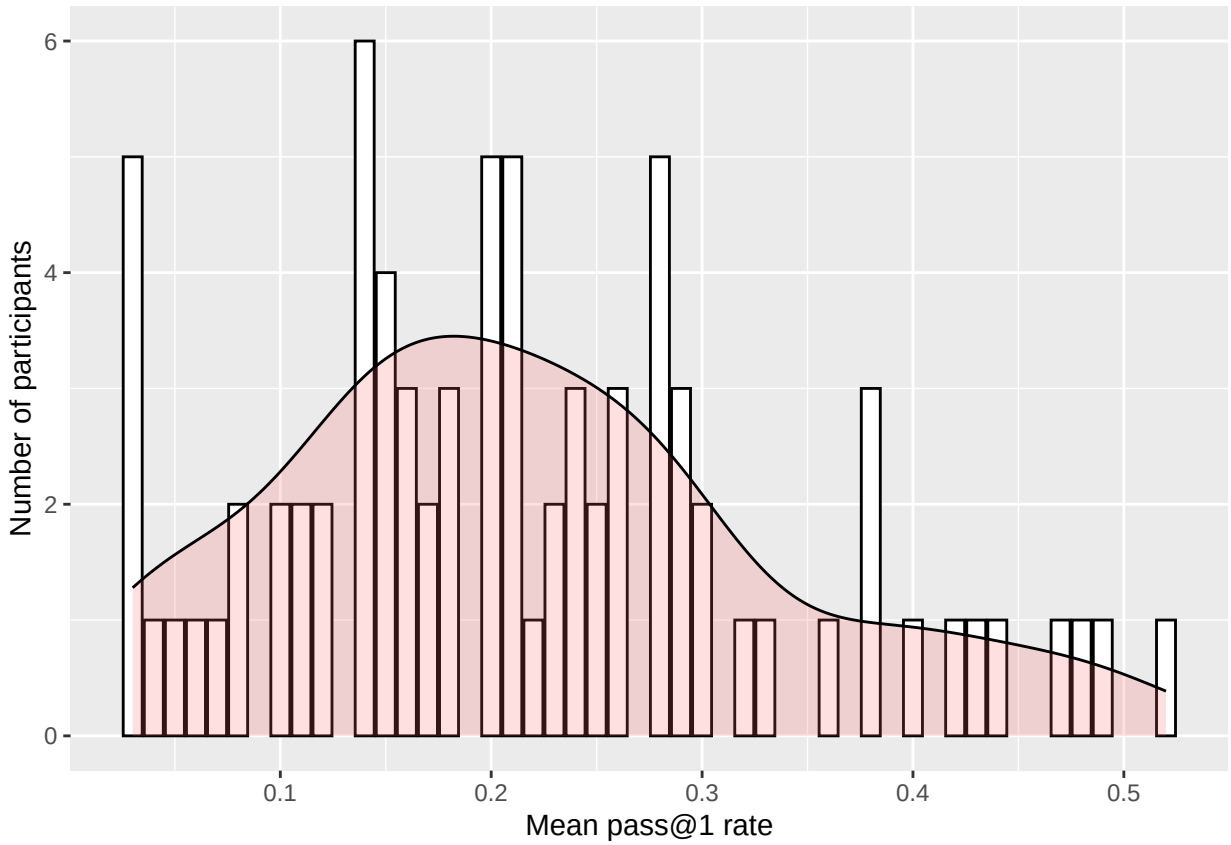
```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```




```
ggsave("n_subs_by_group.pdf",height=10,width=14)
```

```
part.means <- data.all %>% group_by(username) %>% summarize(partMean = round(mean(pass1), digits = 2))
ggplot(data=part.means,aes(x=partMean)) + geom_histogram(stat="count", colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666") + xlab("Mean pass@1 rate") + ylab("Number of participants")
```

```
## Warning in geom_histogram(stat = "count", colour = "black", fill = "white"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```




```
ggsave("participant_success.pdf")
```

```
## Saving 6.5 x 4.5 in image
```