Data Analysis Supplement

EuroVis Short Paper - Submission #1078

7-Apr-2020

This .Rmd contains the analysis and figure generation for:

Crouser, R. Jordan, Ottley, Alvitta, et al. "Investigating the Role of Locus of Control in Moderating Complex Analytic Workflows." 22nd Eurographics Conference on Visualization, 2020.

Import and preprocessing

```
# Interaction data from instrumented system and Google Docs autosave
actions_full <- read.csv("data/eventlog.csv") %>%
  mutate(Time = hms(Time))

# Eliminate incomplete participants and those with data collection errors,
# add more aesthetically-pleasing labels
labels = data.frame("D.ID" = c("VAST5KP", "VASTCER", "VASTFOB",
                               "VASTJL1", "VASTKWO", "VASTL6X",
                               "VASTOIC", "VASTPSZ", "VASTT93",
                               "VASTV8P", "VASTVD5", "VAST6P8",
                               "VASTWCO", "VASTWIO", "VASTWK2",
                               "VAST9L7", "VAST32P", "VASTA6W",
                               "VASTBNJ", "VASTBR1", "VASTBTZ",
                               "VASTBW9"),
                    "D.Name" = c("Participant 1", "Participant 10", "Participant 11",
                                 "Participant 12", "Participant 13", "Participant 14",
                                 "Participant 15", "Participant 16", "Participant 17",
                                 "Participant 18", "Participant 19", "Participant 2",
                                 "Participant 20", "Participant 21", "Participant 22",
                                 "Participant 3", "Participant 4", "Participant 5",
                                 "Participant 6", "Participant 7", "Participant 8",
                                 "Participant 9"))

# Data from qualtrics survey
survey_full <- read.csv("data/survey.csv") %>%
 filter(D.ID %in% labels$D.ID) %>%
  left_join(labels)
```

```
## Joining, by = "D.ID"
```

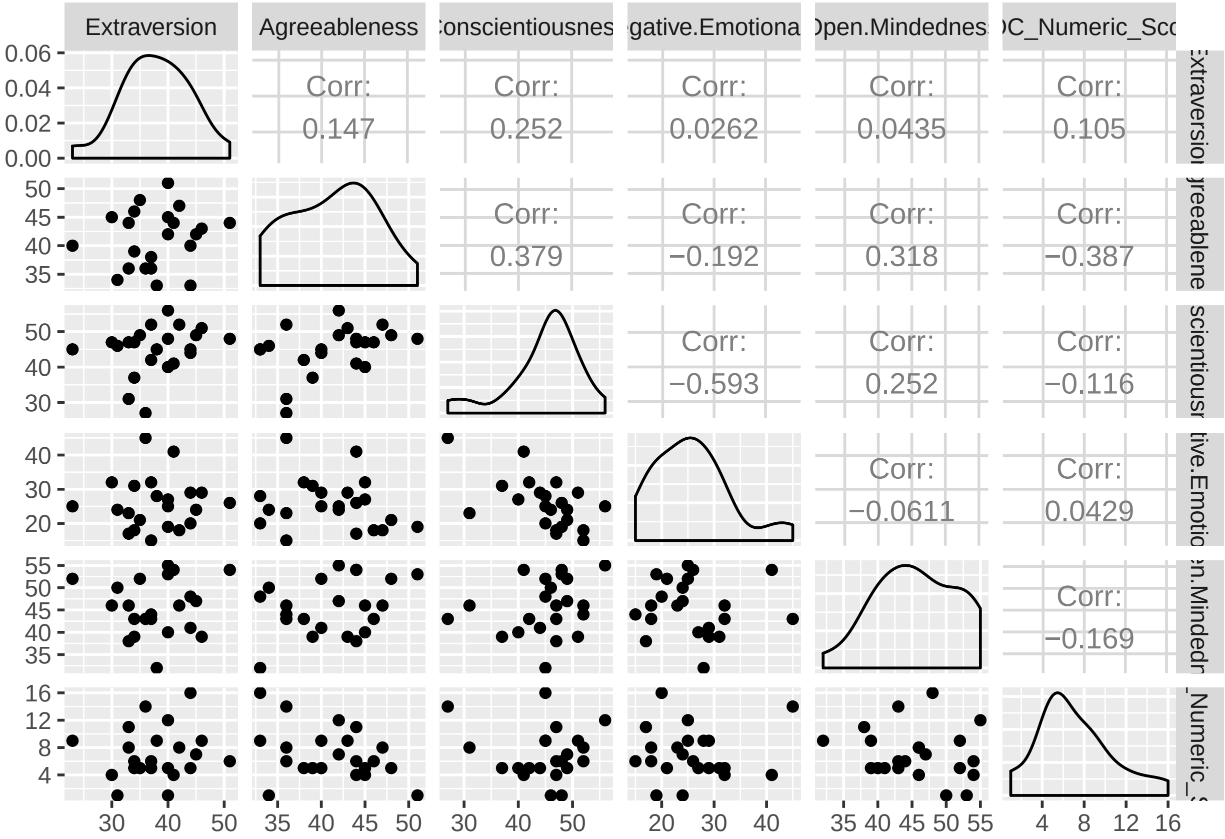
```
# Drop columns for answers to individual questions
survey_summary <- survey_full %>%
  select(-matches('([[:alnum:]]|[[:punct:]])*[[:digit:]]'))
```

Exploration of individual differences

Check for correlation between LOC and 5-Factor Model

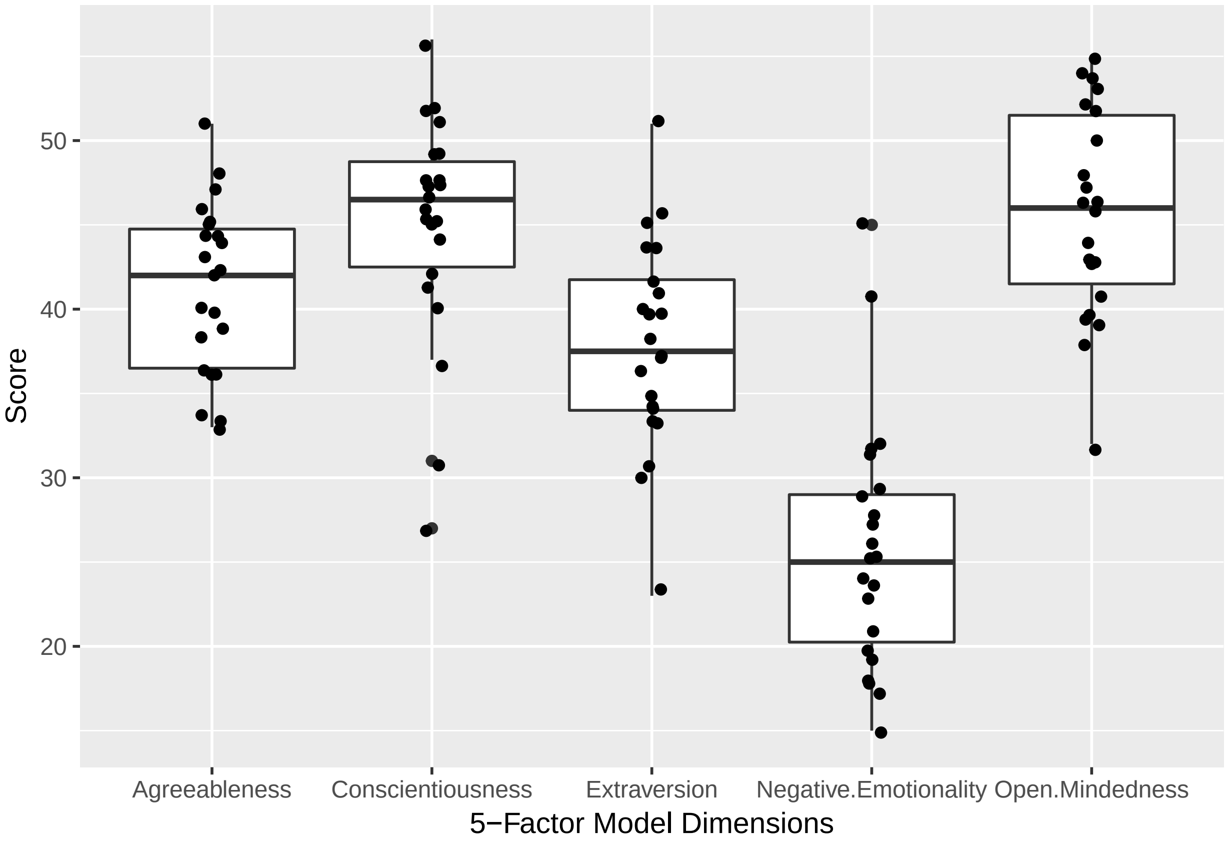
```
individual_differences <- survey_summary %>%
     select(Extraversion:LOC_Numeric_Score)
```

```
ggpairs(individual_differences, progress = F)
```



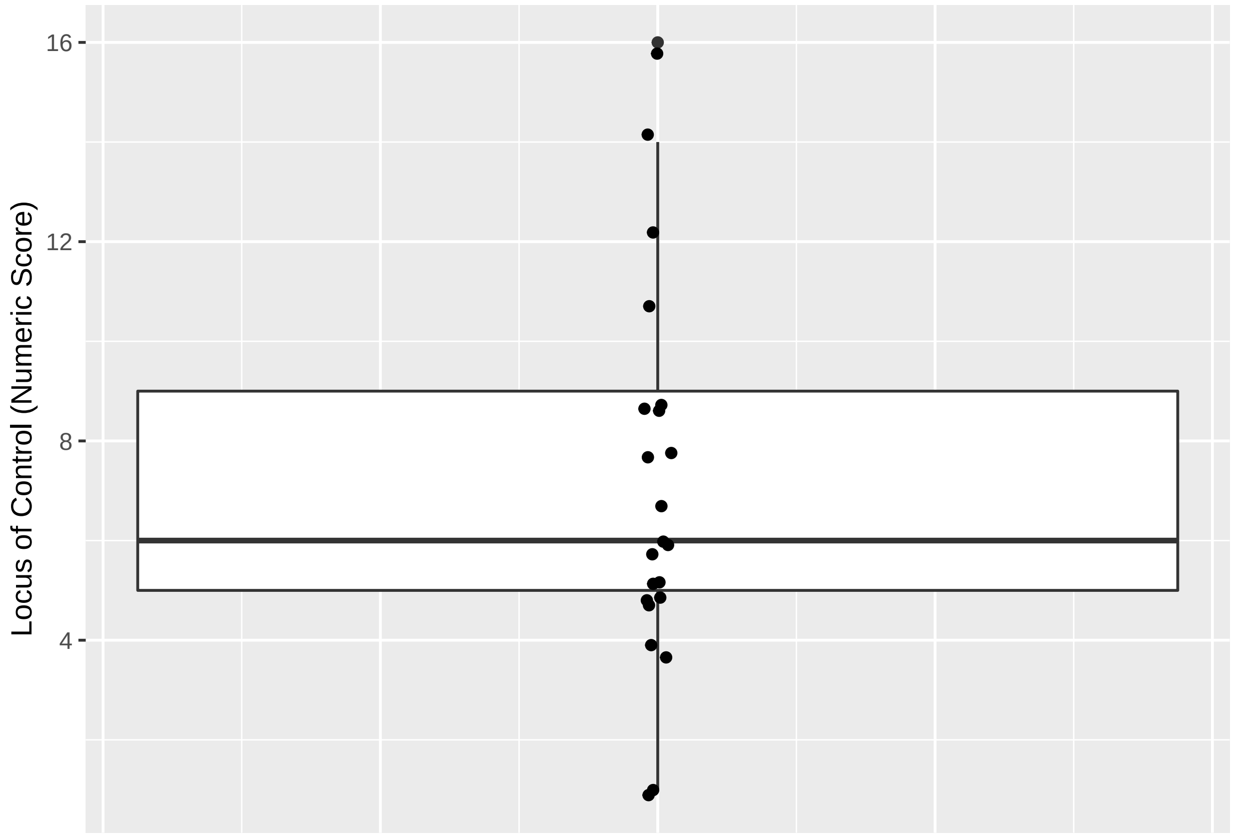
Distribution of participant scores in the 5-Factor Model

```
individual_differences %>%
  gather(key = "attribute", value = "measure", Extraversion:Open.Mindedness) %>%
  ggplot(aes(x=attribute, y=measure)) +
    geom_boxplot() +
    geom_jitter(position=position_jitter(0.05)) +
    xlab("5-Factor Model Dimensions") +
    ylab("Score")
```



Distribution of participants' Locus of Control scores

```
individual_differences %>%
  ggplot(aes(y=LOC_Numeric_Score)) +
    geom_boxplot() +
    geom_jitter(aes(x = 0), position=position_jitter(0.01)) +
    ylab("Locus of Control (Numeric Score)") +
    theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Investigation of overall activity levels

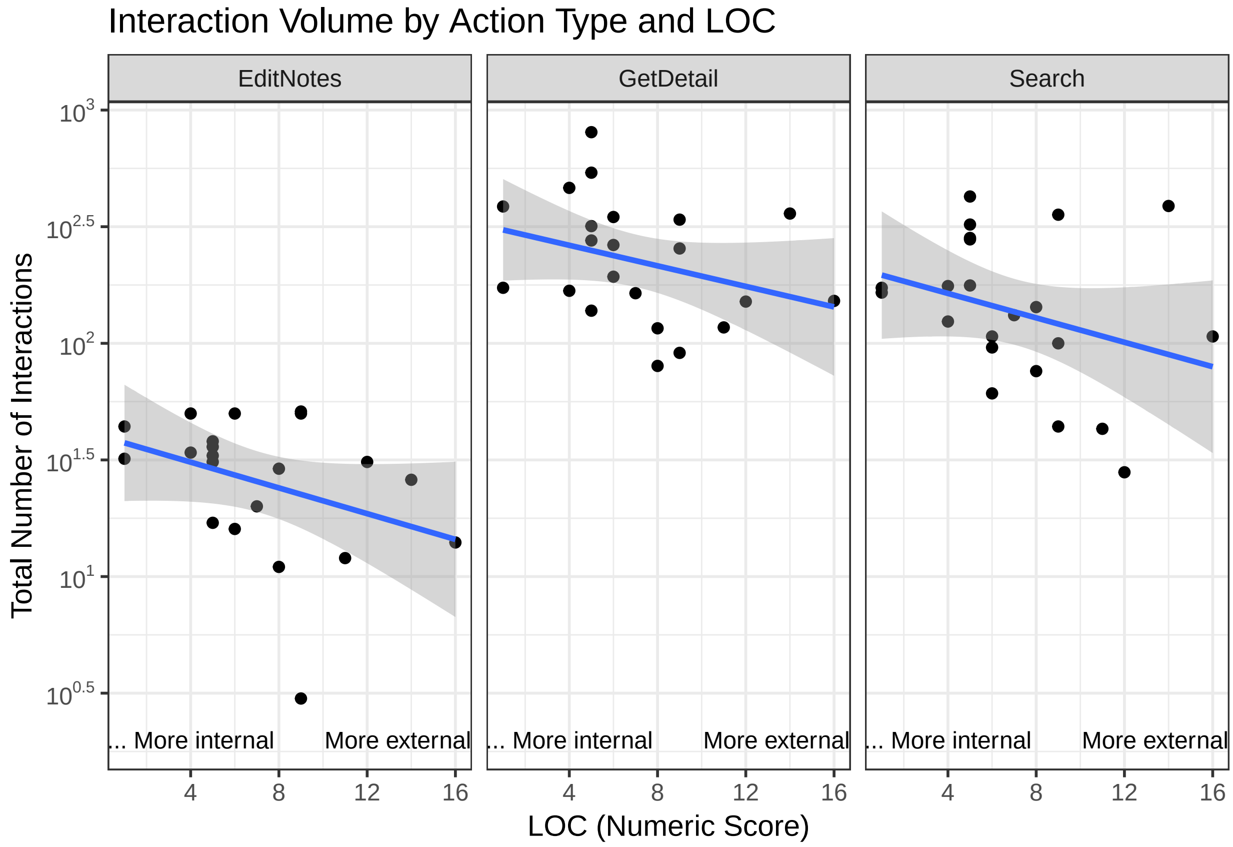
Create aggregate view of activity, totalling up number of actions in each category.

```r
actions_count <- actions_full %>%
  count(Participant, ActionType) %>%
  group_by(Participant) %>%
  mutate(total = sum(n)) %>% # Fill in zeroes
  spread(ActionType, n, fill=0) %>%
  gather(ActionType, n, AddConnection:Search)
```

Visualization of activity levels

Here can see how the total number of actions varies with LOC Numeric Score for each of the three most common action types (`EditNotes`, `GetDetail`, and `Search`):

```r
inner_join(actions_count, survey_summary, by= c("Participant"="D.ID")) %>%
  filter(ActionType != "AddConnection", ActionType != "AddElement") %>%
  filter(n != 0) %>%
  ggplot(aes(x=LOC_Numeric_Score, y=n)) +
  geom_point() +
  geom_smooth(method='lm') +
  ggtitle("Interaction Volume by Action Type and LOC") +
  ylab("Total Number of Interactions") +
  xlab("LOC (Numeric Score)") +
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
            labels = trans_format("log10", math_format(10^.x))) +
  annotate("text", x = 4, y = 2, size = 3, label = "← More internal") +
  annotate("text", x = 14, y = 2, size = 3, label = "More external →") +
  theme_bw() +
  facet_wrap(.~ActionType)
```



We see that for each, those with internal locii of control have the highest activity levels, and those with external LOC scores the lowest levels, respectively.

Investigation of article revisiting behavior

We are interested in examining visiting behaviors. These behaviors are only applicable to the GetDetail action type. The GetDetail action type describes a user choosing to click on an Article, Employee Record, Resume, or Email Header. A revisit is defined as an action of this type wherein the user has previously performed a GetDetail action on that same piece of information.

Data transformation

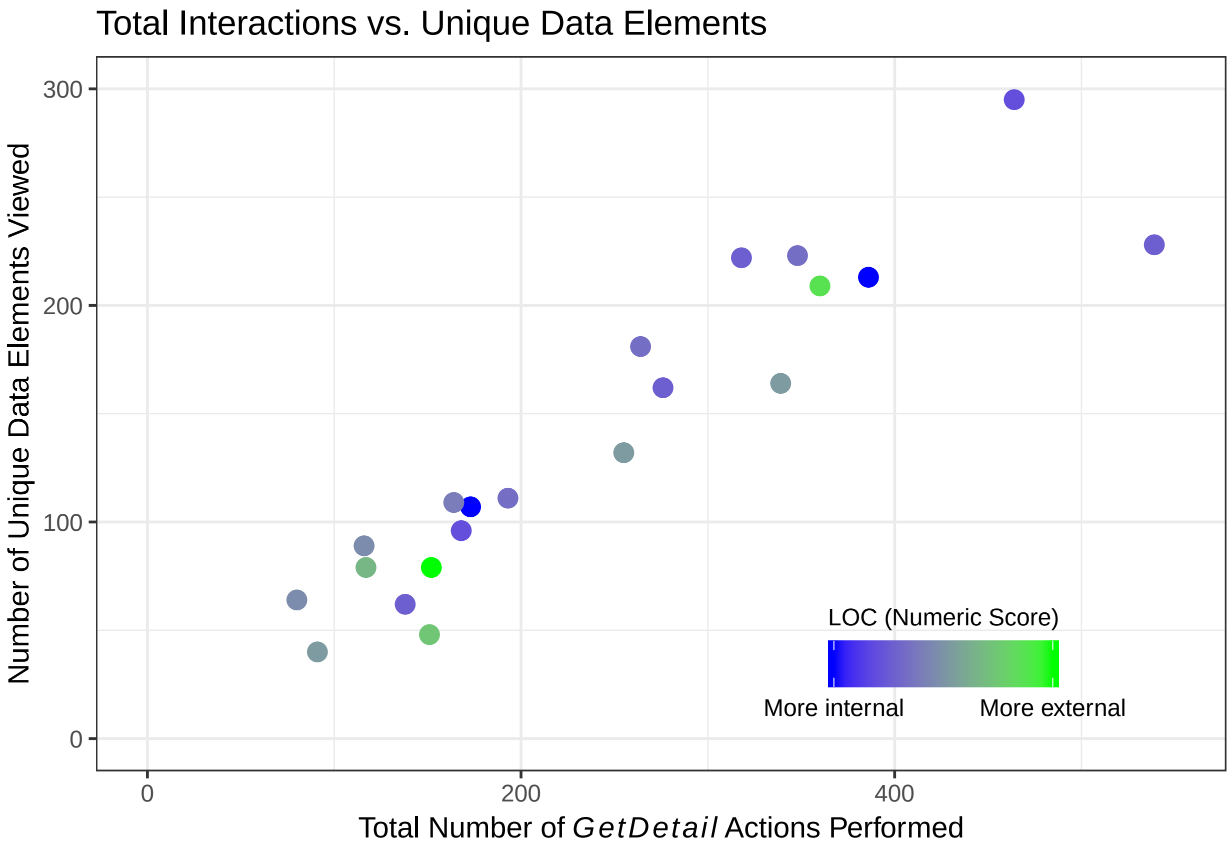
```
# Count revisits
unique_vs_total <- actions_full %>%

  # Count the GetDetail actions for each ActionParameter
  filter(ActionType == "GetDetail") %>%
  group_by(Participant, ActionParameters) %>%
  summarise(n=n()) %>%

  # Aggregate to the participant level, counting unique visits and total visits.
  group_by(Participant) %>%
  summarise(n_unique = n(), n_total=sum(n)) %>%
  # Calculate fraction of unique visits, number of revisits, and join with survey data
  mutate(fraction_unique = (n_unique/n_total)) %>%
  mutate(revisits = (n_total - n_unique)) %>%
  inner_join(survey_full, by= c("Participant"="D.ID"))
```

The above series of transformations delivers an aggregate set of information visited by each participant. We see their total count of GetDetail actions (n_total), the tally of these visits that were visits to a unique piece of data (n_unique), and difference between the two (revisits). We are able to calculate the proportion of GetDetail visits that are visits to a piece of data not previously visited by that participant (fraction_unique).

Visualizing revisit behavior

We now can use the transformed dataset "unique_vs_total" to investigate revisits.

```
ggplot(unique_vs_total, aes(x=n_total,y=n_unique)) +
  geom_point(aes(color=LOC_Numeric_Score), size = 3) +
  ggtitle("Total Interactions vs. Unique Data Elements") +
  ylab("Number of Unique Data Elements Viewed") +
  xlab("Total Number of"~italic(GetDetail)~"Actions Performed") +
  scale_colour_gradient(name = "LOC (Numeric Score)",
                        low = "blue", high = "green",
                        breaks=c(1,16),
                        labels=c("More internal","More external")) +
  theme_bw() +
  theme(legend.direction = "horizontal",
        legend.position = c(0.75, 0.15),
        legend.background = element_rect(fill="transparent"),
        legend.title.align=0.5,
        legend.title = element_text(size = 9),) +
  guides(colour = guide_colourbar(title.position="top")) +
  xlim(c(0,550)) +
  ylim(c(0,300))
```



The above shows a strong positive linear correlation between the number of unique visits and the total number of visits by each participant.

Coloring the points by LOC category shows us little except for what we already knew about the differences in activity level between the LOC categories.

```r
unique_vs_total %>%
  lm(formula = n_unique ~ n_total) %>%
  summary()
```

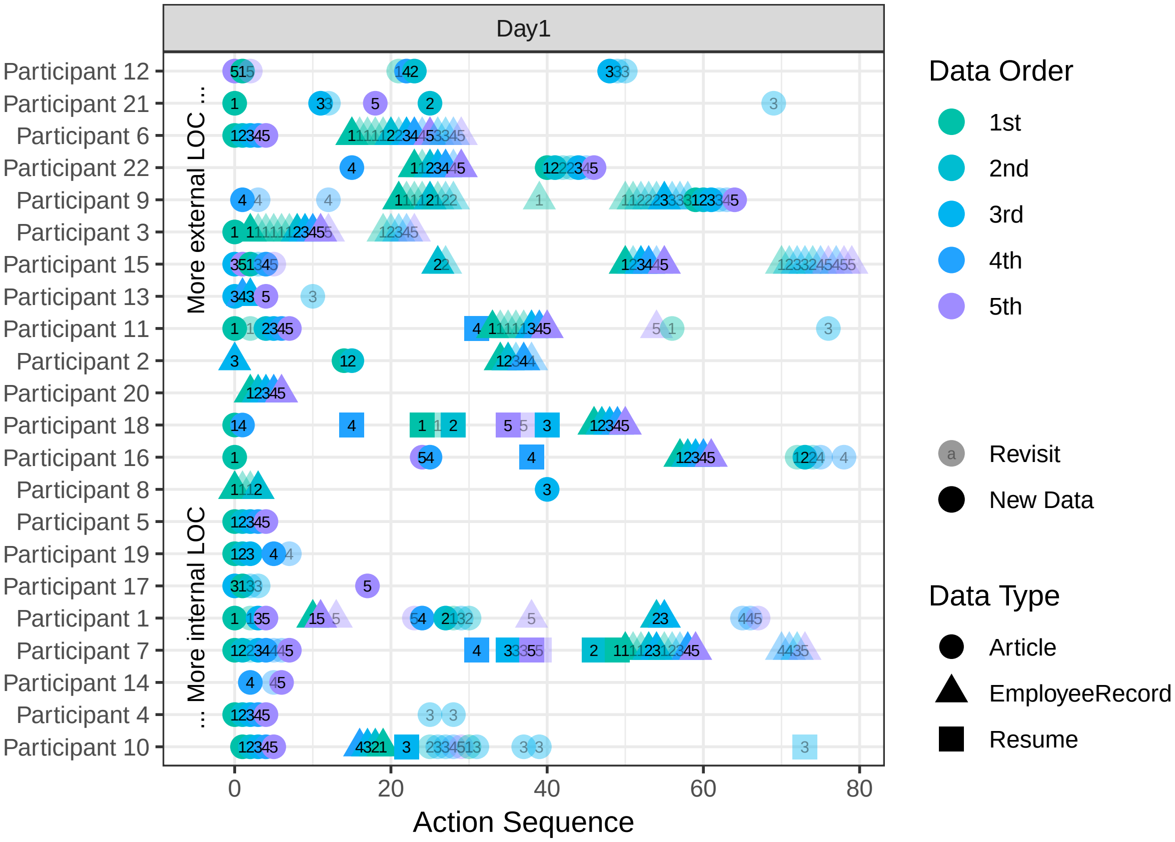
```
##
## Call:
## lm(formula = n_unique ~ n_total, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.094 -13.464   4.325  16.065  40.825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.74269   10.80664   0.531    0.601
## n_total      0.55167    0.03411  16.172 5.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.03 on 20 degrees of freedom
## Multiple R-squared:  0.929,  Adjusted R-squared:  0.9254
## F-statistic: 261.5 on 1 and 20 DF,  p-value: 5.961e-13
```

A linear mode confirms what we see in the visualization: `n_total` is a significant predictor of `n_unique`, explaining 93% of the variability in number of unique items visited.

Anchoring behavior and "clicking-down" meta actions on the first 5 Articles, Employee Records, and Resumes

Here we examine visits to the first five data items of each type that appear in the list as presented to the participants.

```r
# Label revisits
revisits_over_time <- actions_full %>%
  filter(ActionType == "GetDetail") %>%
  arrange(Participant, Date, period_to_seconds(hms(Time))) %>%
  group_by(Participant, ActionParameters) %>%
  mutate(id = row_number() -1) %>%
  mutate(revisit = ifelse(id > 0, TRUE, FALSE)) %>%
  inner_join(survey_summary, by= c("Participant"="D.ID")) %>%
  group_by(Participant) %>%
  mutate(index = row_number()-1) %>%
  ungroup() %>%
  group_by(Participant, Day) %>%
  mutate(action_order_by_day = row_number()-1)

# Note: this subset only contains actions of the GetDetail type.
# Any action between successive GetDetails is not counted.

# Create list of first 5 items, in order of appearances for all Article, Resume,
# and EmployeeRecord InfoType
info_classes <- data.frame("Name" = c("Article 406",
                                      "Article 121",
                                      "Article 265",
                                      "Article 227",
                                      "Article 56",
                                      "Resume-IsandeBorrasca.pdf",
                                      "Resume-LidelseDedos.pdf",
                                      "Resume-SvenFlecha.pdf",
                                      "Resume-CorneliaLais.pdf",
                                      "Resume-NilsCalixto.pdf",
                                      "Employee Record 1",
                                      "Employee Record 2",
                                      "Employee Record 3",
                                      "Employee Record 4",
                                      "Employee Record 5"),
                           "Order" = factor(rep(1:5, length.out = 15)),
                           "InfoType" = factor(rep(c("Article",
                                                     "Resume",
                                                     "EmployeeRecord"),
                                                   each = 5)))

revisits_over_time %>%
  filter(ActionParameters %in% info_classes$Name,
         Day == "Day1",
         action_order_by_day < 80) %>%
  inner_join(info_classes, by=c("ActionParameters"="Name")) %>%
  ggplot(aes(x = action_order_by_day,
          y = reorder(D.Name, LOC_Numeric_Score),
          alpha=!revisit, col=Order, pch=InfoType)) +
  geom_point(size = 4) +
  geom_text(aes(label=Order), size = 2, color = "black")+
  facet_grid(.~Day) +
  scale_alpha_discrete(range=c(0.4, 1), labels = c("Revisit", "New Data")) +
  scale_colour_hue(h = c(180, 270), labels = c("1st", "2nd", "3rd", "4th", "5th")) +
  ylab("") +
  xlab("Action Sequence") +
  annotate("text", x = -5, y = 5, size = 3, label = "← More internal LOC" , angle = 90)+
  annotate("text", x = -5, y = 18, size = 3, label = "More external LOC →", angle = 90) +
  theme_bw() +
  labs(color = "Data Order",
       alpha = "",
       pch = "Data Type")
```


Day1
Participant 12
Participant 21
Participant 6
Participant 22
Participant 9
Participant 3
Participant 15
Participant 13
Participant 11
Participant 2
Participant 20
Participant 18
Participant 16
Participant 8
Participant 5
Participant 19
Participant 17
Participant 1
Participant 7
Participant 14
Participant 4
Participant 10
More external LOC ...
... More internal LOC
Action Sequence
Data Order
1st
2nd
3rd
4th
5th
Revisit
New Data
Data Type
Article
EmployeeRecord
Resume