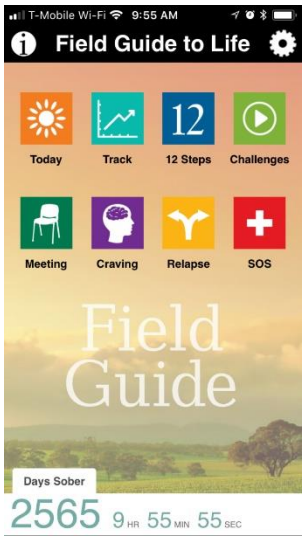
Author: Benji Fischman

Data-driven Support of Addiction Recovery Communities via Classification and Regression Analysis.

Introduction How might we inform the addiction recovery community on how to improve the probability of a successful recovery? Substance addiction is a public health epidemic in America. This project aims to assist those suffering from substance addiction so that they might better understand their demons. Visualizations help communicate high-level takeaways to those unfamiliar with statistics and provides a sanity check for the prediction model. Tenfold-cross validation shows the model is 84% accurate in a classification context. A simple survey assists users unfamiliar with a structured query language explore the dataset. The visual display of quantitative information is a non-trivial task, but a mixed methods approach to problem-solving help make temporal sequence analysis tractable and pragmatic.

Dataset The initial dataset consisted of three years' worth of action history for 3346 distinct users and was provided by the Field Guide to Life mobile application [12]. In total, the original dataset had 407,806 observations. Users could perform 32 unique actions while using the Field Guide to Life application (e.g., view a message of the day, self-report a relapse, view a sober day counter, etc.). Formative qualitative analysis with recovering addicts suggests that aggregation of actions into seven categories is a suitable heuristic for the domain of interest. The seven categories include meetings, relapse, referencing any of the 12 steps, tracking progress, seeking support, opening the application, and interacting with the message of the day.



Screenshot from Field Guide to Life

	App.Instance.ID	App.Name	Platform	Data.Point.Name	Data.Point.Code	Timestamp
220622	2885f0f40fab492da0441a2b25500478	Mobile MORE Field Guide to Life Container	Google	Meeting	MEETING_AA	2015-03-18 03:55:27

Example Observation Data Point

Exclusion Criteria Clean data is essential to any statistical analysis, so data preparation should exclude any superfluous noise in the dataset. Any observations that occur after 2016-10-26 were considered an error (e.g., the date 2557-08-18 has not happened yet so that observation is an error). Additionally, users that used the app for less than thirty days were excluded, and individuals that only open the application with no other actions were considered out of scope. In total 87 users and 5527 observations were removed (i.e., N=402,648 and user count = 3259).

Prediction Model This work is at the intersection of two domains: addiction recovery, and mobile health app data analytics. Prior work in both areas helps focus the ad-hoc creation of the prediction model. In this section, I will briefly summarize the predictor variable, and factors used

in the ANOVA regression and classification models. Literature suggests temporality is one of the most salient factors in drinking habits of recovering alcoholics, so abstraction of this idea to substance addiction is a reasonable heuristic [9]. To leverage temporality as a meaningful variable user session identification was applied based on strong regularities in inter-activity time [3]. For instance, if no actions were performed within an hour of the previous user action we end that users' session. For each user, the mean of the cumulative count of relapses divided by the cumulative count of sessions represents the probability of relapse. Feature selection for the predictor variable was an iterative process. Factors that influence the predictor variable include: engagement (i.e., mean of the cumulative sum of actions sans relapse reports divided by cumulative sum of active days of app use); action diversity (i.e., count of action performed sans relapse divided by count of total possible actions); time interval of initial app usage till a user's' first reported relapse, and the mean of the cumulative sum of daily message views divided by the cumulative sum of total action sans relapse. All statistical analysis was conducted in R (version 3.3.3), using packages that include: dplyr, zoo, and data.tables [1,8,10]. Finally, normalization of engagement and daily message view was applied such that the variables range between 0 and 1.

Model Validation A sanity check of the proposed model, is essential and necessary before investing time to develop an interactive visualization that communicates the high-level takeaways. Classification of which users will relapse is a suitable approach to experimental model validation[4,5]. Ten-fold cross-validation of the model was 84% accurate, which is good enough to argue the proposed model is reliable and valid. Ten-fold cross-validation is the process of randomly splitting the dataset into ten partitions, training the model to classify on nine of the ten partitions, and then measuring the errors of classification in the tenth test partition. The R package rpart was used for training the classification and ANOVA regression models [6].

Visualization Clear visual communication of quantitative information is crucial when catering to a community that likely has a limited grasp of statistics [7]. Two novel visualization techniques are leveraged: temporal alignment of event sequences and patterns using EventFlow, and a custom interactive decision tree using d3 [2,11]. The project aims to help recovering addicts, so to better extract signals that might elucidate how to improve an individual's chances of a successful recovery I exclude all users that never report a relapse. This drastically reduces the size of the dataset (N= 48,682, user count = 591). The default settings in rpart were used, except for the minimum number of observations (30) in a node to compute a split as well as setting the complexity parameter threshold of a split to the minimum mean error squared found using the ten-fold cross-validation (0.0157). To prune the tree at the optimal depth, I examined the complexity parameter plot of the tree vs. the relative confusion error at each depth (Figure 1). I then pruned the tree to the level which minimized relative error.

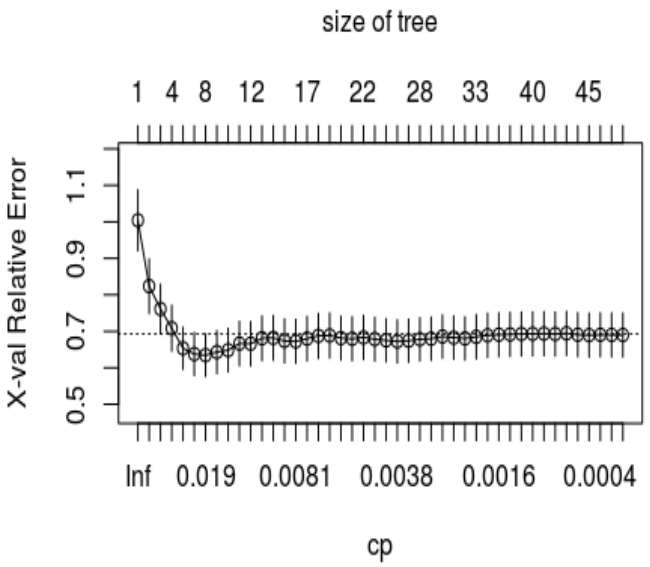


Figure 1

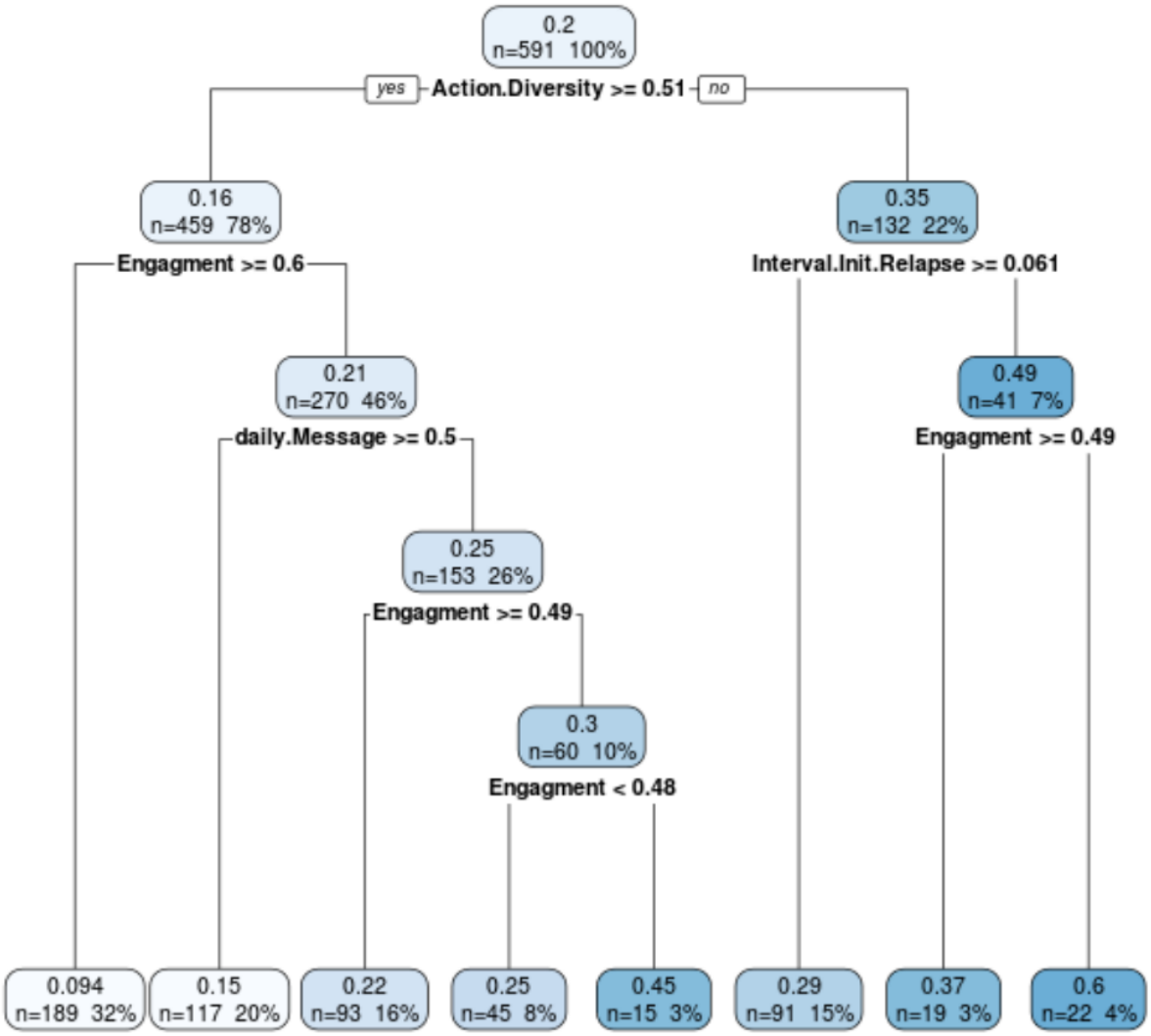
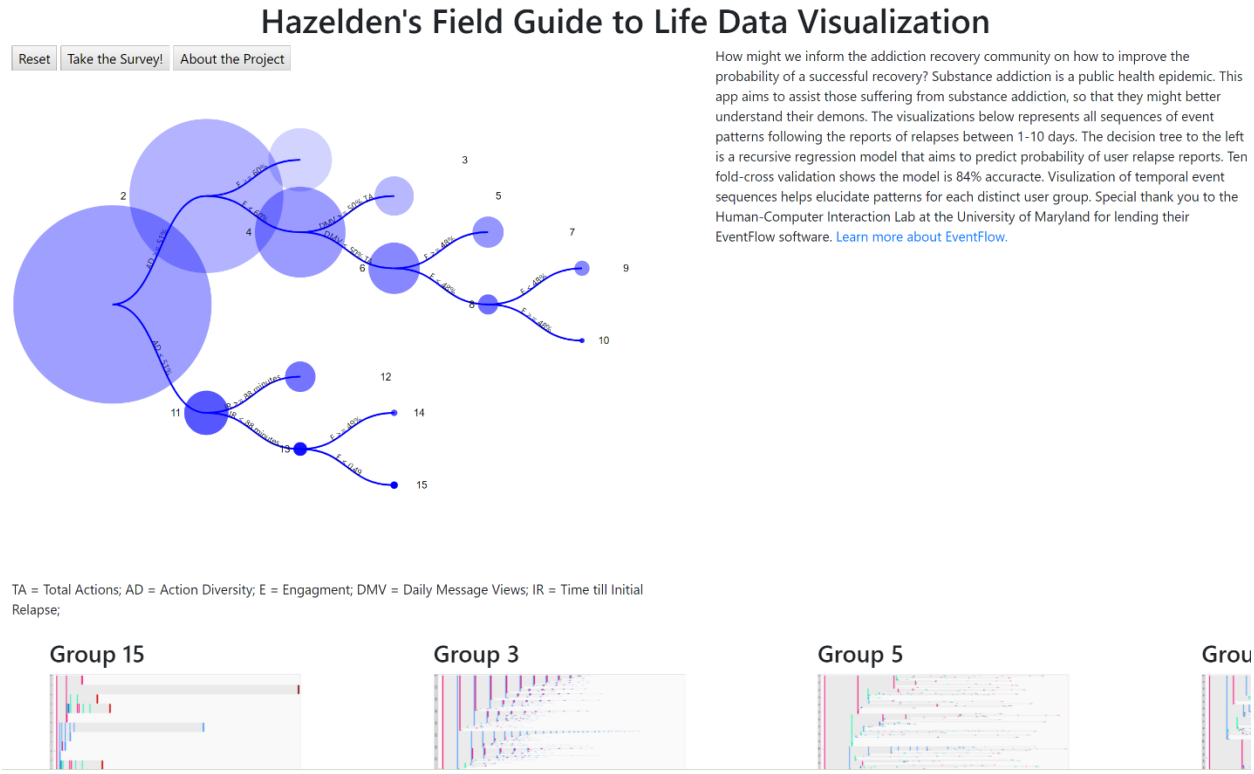


Figure 2 (rpart output)

Figure 2 represents an ANOVA regression represented as a decision tree, which is output from rpart. Figure 2 was a good start to the data visualization, but the visualization still needed interactivity to foster data exploration. To maximize the impact and reach of this project I wanted anyone to be able to access the visualizations through a browser, and querying the dataset had to be as simple as taking a survey. Below is a screenshot of the web app I implemented using expressJS and d3 [11,13].



Web App 1

The radius of each circle in Web App 1 corresponds to the number of users that fall into a distinct cohort of users. The labels along the edges denote predicates each user must satisfy to fall into a given subset. The opacity of each circle signifies the mean probability of relapse for each user in each group normalized between 0 and 1. For example, the darker the circles

Welcome to the Hazelden Data Exploration App!

Please complete the form below to discover how to improve your odds of a staying clean and sober!

How long have you been involved in a 12 step recovery program?

Period: Years ▼ | Duration: 3

How often do you attend 12 step recovery meetings?

Period: Weekly ▼ | Frequency: 1

How often do you seek support during your recovery process?

Period: Monthly ▼ | Frequency: 1

How often have you kept track of progress?
(e.g. Make a journal entry, kept a sober counter, etc.)

Period: Monthly ▼ | Frequency: 2

How often do you read inspiriational messages/message of the day?

Period: Daily ▼ | Frequency: 1

How often do you reference of the 12 steps?

Period: Monthly ▼ | Frequency: 1

Have you ever relapsed while in a 12 step recovery program?

- (•) Yes
- ( ) No

If you have relapsed, did you remain sober for at least one day after you first started attending 12 step meetings?
(i.e. Duration until initial relapse greater than one day.)

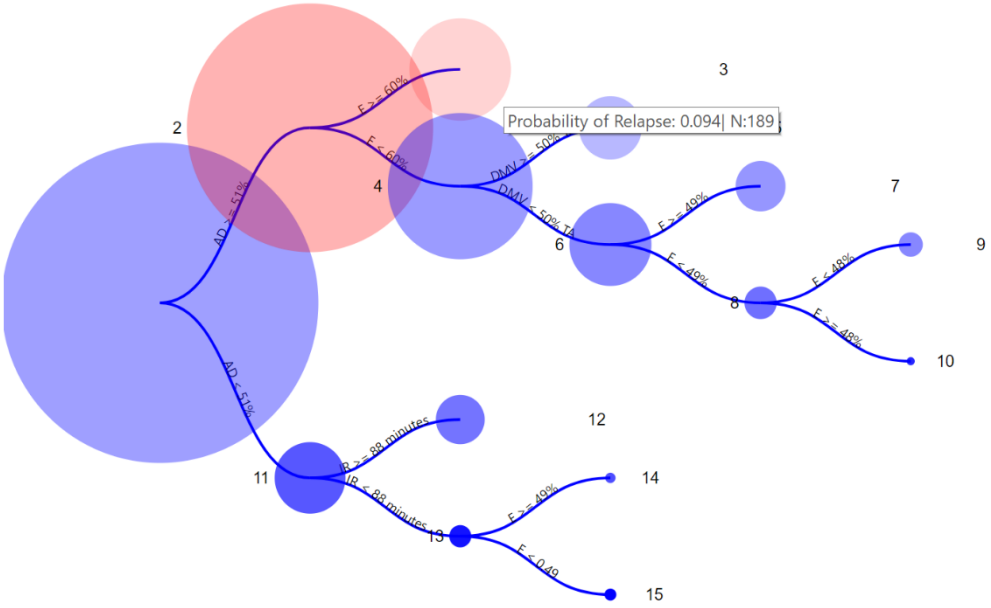
- ( ) Yes
- (•) No

Please respond to the prompt, "I am engaged in my recovery proccess."

Strongly agree | (•) Agree | Neutral | Disagree | Strongly disagree

Web App 2

represent users that are more likely to relapse, and the lighter circles represent the users that have a better chance of a staying clean and sober. Additionally, tooltips appear on hover that explicitly state the probability and number of users in each group. The numbers are labels for each of the user groups, which help the user connect the second aspect of data visualization (Figure 2) to the decision tree. Additionally, the decision tree has pan and zoom functionality. The survey feature seen in Web App 2 helps guide users to the group that is most like their recovery process by highlighting the groups they fall within in red.



Web App 3

The user can then choose to view a static visualization of any of the leaf node groups. On click of one of the static image thumbnails seen at the bottom of Web App 1, the user is directed to a robust visualization of temporal event sequences and patterns aligned for each relapse report of 1-10 days. I excluded daily message views from the visualization to decrease the noise in the visual inspection of pattern recognition [2].

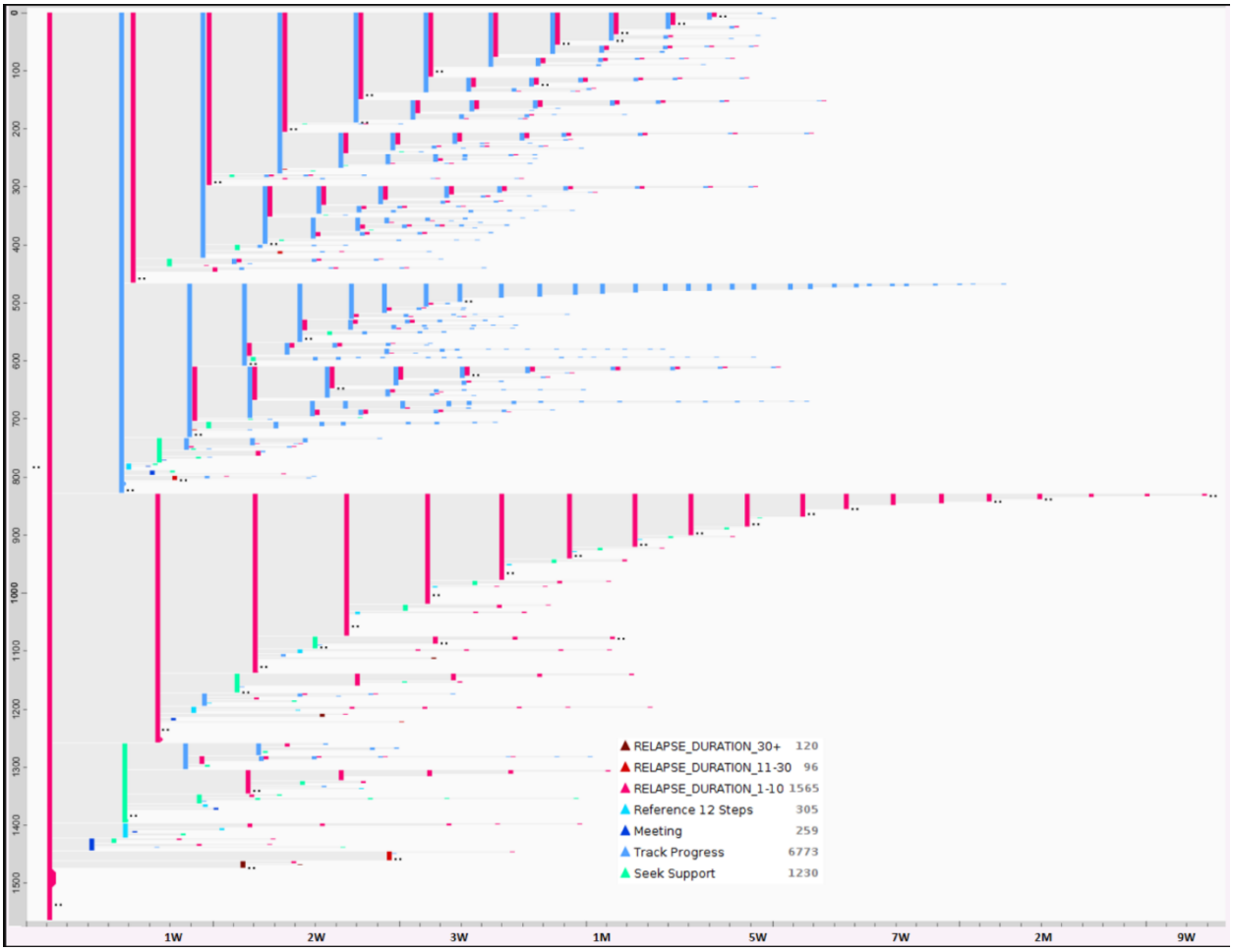


Figure 2 (Group 3)

Figure 2's interpretation is as follows, the x-axis is time, and the y-axis is every sequence of events following the report of a relapse in descending rank of the number of events following the relapse report. More information on interpretation of EventFlow visualizations can be found elsewhere [2].

Findings & Conclusion

Figure 2 is group 3, which had the lowest probability of relapse. Notice the long blue line spanning from 0 to around 800; this line means a tracking progress event followed 800 relapse reports. Also note the length of the second red line that begins around 800 and ends around 1250, which decreases over time at intervals of about one week. This trend might suggest that weekends can be a trigger for many, but tracking progress consistently will help increase a user's engagement. At a high level, the data suggests that users are more likely to stay clean and sober if they are honest with the twelve-step recovery process (i.e., report cravings prior to a relapse, report the relapse, and reset the sobriety counter after a relapse). Additionally, 41 users report a relapse within 88 minutes of first using the app, which may imply user motivations for app usage are correlated with relapse cravings, but a casual statistical analysis remains as future work.

Visualization of each of the groups provides an additional sanity check for the ad-hoc prediction model. For example, on an earlier iteration, the EventFlow visualization helped point out that temporality (i.e., user session identification) is crucial to the facilitation of a normal distribution of events both between and within groups. A simple visual inspection of the eight distinct cohorts confirms relatively normal distributions of event types [3]. Informal evaluation with real recovering addicts has gotten very positive feedback!

Citations

1. Matt Dowle, Arun Srinivasan, Jan Gorecki, Tom Short, Steve Lianoglou, and Eduard Antonyan. 2017. *data.table: Extension of "data.frame."* Retrieved from https://cran.r-project.org/web/packages/data.table/index.html
2. Fan Du, Ben Shneiderman, Catherine Plaisant, Sana Malik, and Adam Perer. 2017. Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus. *IEEE Transactions on Visualization and Computer Graphics* 23, 6: 1636–1649. https://doi.org/10.1109/TVCG.2016.2539960
3. Aaron Halfaker, Oliver Keyes, Daniel Kluver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, and Morten Warncke-Wang. 2015. User Session Identification Based on Strong Regularities in Inter-activity Time. 410–418. https://doi.org/10.1145/2736277.2741117
4. Stephenie C. Lemon, Jason Roy, Melissa A. Clark, Peter D. Friedmann, and William Rakowski. 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine* 26, 3: 172–181.
5. Katrina J. Serrano, Kisha I. Coa, Mandi Yu, Dana L. Wolff-Hughes, and Audie A. Atienza. 2017. Characterizing user engagement with health app data: a data mining approach. *Translational Behavioral Medicine* 7, 2: 277–285. https://doi.org/10.1007/s13142-017-0508-y
6. Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley. 2017. Package "rpart." *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.
7. Edward R. Tufte. 1986. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.
8. Hadley Wickham, Romain Francois, Lionel Henry, Kirill Müller, and RStudio. 2017. *dplyr: A Grammar of Data Manipulation*. Retrieved from https://cran.r-project.org/web/packages/dplyr/index.html
9. Katie Witkiewitz and Katherine E. Masyn. 2008. Drinking trajectories following an initial lapse. *Psychology of Addictive Behaviors* 22, 2: 157–167. https://doi.org/10.1037/0893-164X.22.2.157
10. Achim Zeileis, Gabor Grothendieck, Jeffrey A. Ryan, Joshua M. Ulrich, and Felix Andrews. 2017. *zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations)*. Retrieved from https://cran.r-project.org/web/packages/zoo/index.html
11. *d3: Bring data to life with SVG, Canvas and HTML*. Retrieved from https://github.com/d3/d3
12. Mobile More Field Guide to Life App -- Hazelden. Retrieved November 18, 2017 from https://www.hazelden.org/web/go/fieldguide
13. Express - Node.js web application framework. Retrieved November 18, 2017 from https://expressjs.com/