Term Project

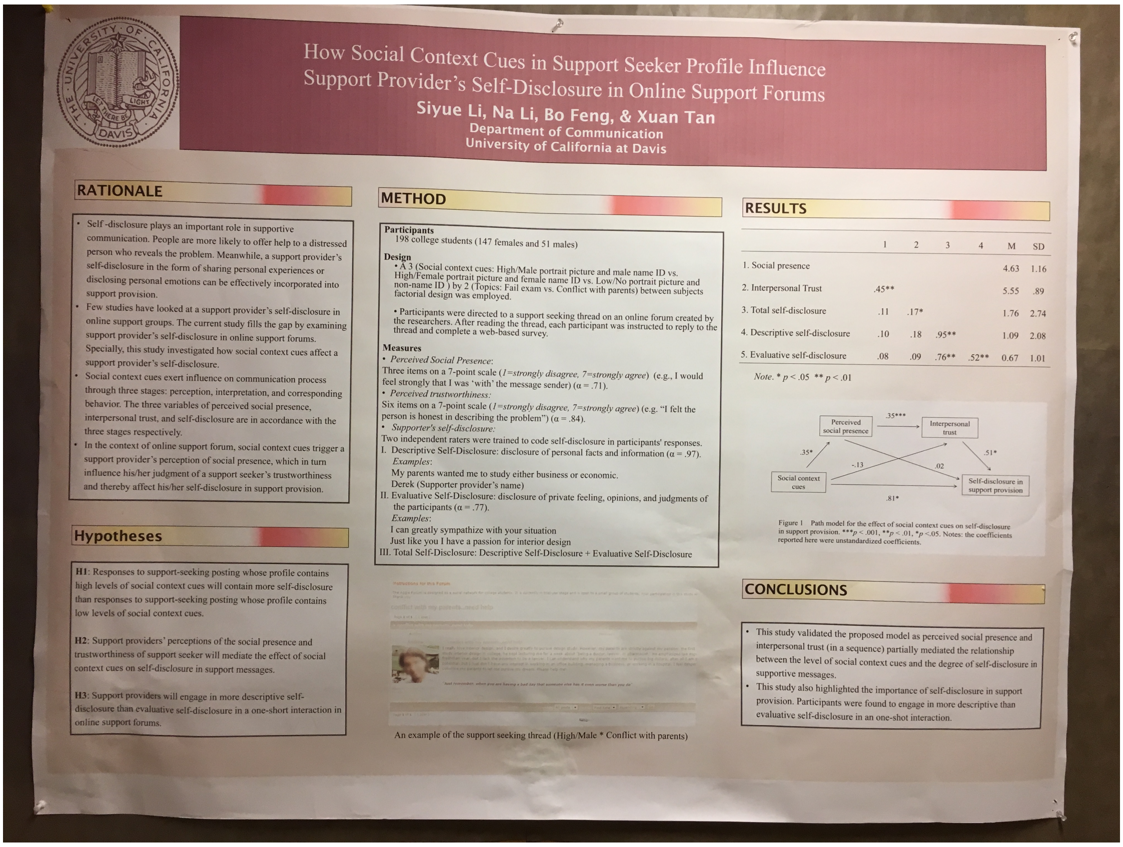
Lynn Miyashita, Alisha Nanda, Sabrina Sheu, Sam Waters

March 21, 2019

1. Problem A

1. Lynn Miyashita

   The research report *How Social Context Cues in Support Seeker Profile Influence Support Provider's Self-Disclosure in Online Support Forums* shows the relationship between amount of social context cues and the response level, as well as how self-disclosure affects the online support forum presence of the support provider. Siyue Li, Na Li, Bo Feng, and Xuan Tan took three hypotheses and researched using a methodology that varied in the amount of social context cues were provided. This showed in the amount of information provided on the online forum for things such as gender, profile picture, or some sort of ID for their name.



How Social Context Cues in Support Seeker Profile Influence Support Provider's Self-Disclosure in Online Support Forums

Siyue Li, Na Li, Bo Feng, & Xuan Tan
Department of Communication
University of California at Davis

RATIONALE

- Self -disclosure plays an important role in supportive communication. People are more likely to offer help to a distressed person who reveals the problem. Meanwhile, a support provider's self-disclosure in the form of sharing personal experiences or disclosing personal emotions can be effectively incorporated into support provision.
- Few studies have looked at a support provider's self-disclosure in online support groups. The current study fills the gap by examining support provider's self-disclosure in online support forums. Specially, this study investigated how social context cues affect a support provider's self-disclosure.
- Social context cues exert influence on communication process through three stages: perception, interpretation, and corresponding behavior. The three variables of perceived social presence, interpersonal trust, and self-disclosure are in accordance with the three stages respectively.
- In the context of online support forum, social context cues trigger a support provider's perception of social presence, which in turn influence his/her judgment of a support seeker's trustworthiness and thereby affect his/her self-disclosure in support provision.

Hypotheses

H1: Responses to support-seeking posting whose profile contains high levels of social context cues will contain more self-disclosure than responses to support-seeking posting whose profile contains low levels of social context cues.

H2: Support providers' perceptions of the social presence and trustworthiness of support seeker will mediate the effect of social context cues on self-disclosure in support messages.

H3: Support providers will engage in more descriptive self-disclosure than evaluative self-disclosure in a one-short interaction in online support forums.

METHOD

Participants
198 college students (147 females and 51 males)

Design
- A 3 (Social context cues: High/Male portrait picture and male name ID vs. High/Female portrait picture and female name ID vs. Low/No portrait picture and non-name ID) by 2 (Topics: Fail exam vs. Conflict with parents) between subjects factorial design was employed.
- Participants were directed to a support seeking thread on an online forum created by the researchers. After reading the thread, each participant was instructed to reply to the thread and complete a web-based survey.

Measures
- *Perceived Social Presence*:

Three items on a 7-point scale (*1=strongly disagree, 7=strongly agree*) (e.g., I would feel strongly that I was 'with' the message sender) (α = .71).
- *Perceived trustworthiness:*

Six items on a 7-point scale (*1=strongly disagree, 7=strongly agree*) (e.g. "I felt the person is honest in describing the problem") (α = .84).
- *Supporter's self-disclosure:*

Two independent raters were trained to code self-disclosure in participants' responses.

I. Descriptive Self-Disclosure: disclosure of personal facts and information (α = .97).
Examples:
My parents wanted me to study either business or economic.
Derek (Supporter provider's name)
II. Evaluative Self-Disclosure: disclosure of private feeling, opinions, and judgments of the participants (α = .77).
Examples:
I can greatly sympathize with your situation
Just like you I have a passion for interior design
III. Total Self-Disclosure: Descriptive Self-Disclosure + Evaluative Self-Disclosure

Instructions for this Forum
conflict with my parents...need help
"Just remember, when you are having a bad day that someone else has it even worse than you do"

An example of the support seeking thread (High/Male * Conflict with parents)

RESULTS

	1	2	3	4	M	SD
1. Social presence					4.63	1.16
2. Interpersonal Trust	.45**				5.55	.89
3. Total self-disclosure	.11	.17*			1.76	2.74
4. Descriptive self-disclosure	.10	.18	.95**		1.09	2.08
5. Evaluative self-disclosure	.08	.09	.76**	.52**	0.67	1.01

Note. * $p < .05$ ** $p < .01$

Figure 1 Path model for the effect of social context cues on self-disclosure in support provision. ***$p < .001$, **$p < .01$, *$p < .05$. Notes: the coefficients reported here were unstandardized coefficients.

CONCLUSIONS

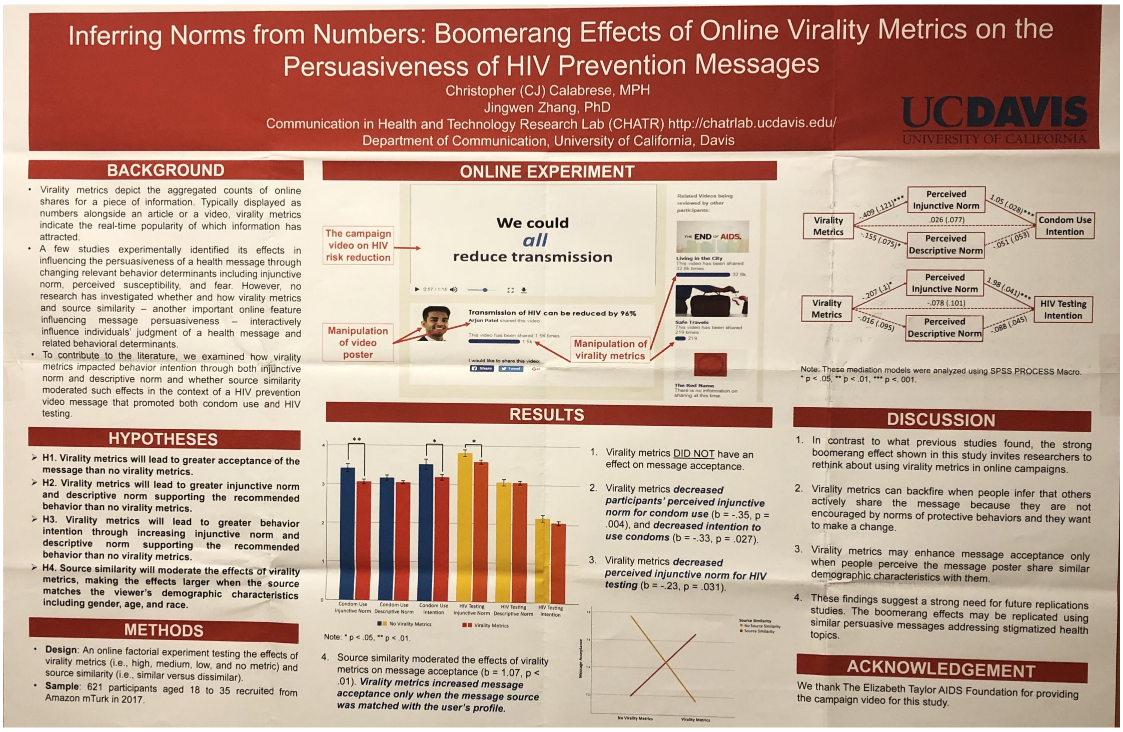
- This study validated the proposed model as perceived social presence and interpersonal trust (in a sequence) partially mediated the relationship between the level of social context cues and the degree of self-disclosure in supportive messages.
- This study also highlighted the importance of self-disclosure in support provision. Participants were found to engage in more descriptive than evaluative self-disclosure in an one-shot interaction.

2. Alisha Nanda

3. Sabrina Sheu

4. Sam Waters

In their research report, *Inferring Norms from Numbers: Boomerang Effects of Online Virality Metrics on the Persuasiveness of HIV Prevention Messages*, Christopher Calabrese, MPH, and Jingwen Zhang, PhD, discuss the effectiveness of virality metrics on the effectiveness of an online campaign's message. "Virality metrics" are more colloquially known as the statistics shown alongside online videos, such as its number of views, shares, and likes. Calabrese and Zhang specifically researched the effects of an HIV prevention video's virality metrics on both injunctive and descriptive norm. According to the 1990 paper, *A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places*, descriptive norms are typically accepted patterns of societal behavior; They are the "what is", while injunctive norms are behaviors which are perceived to be accepted by others – the "what ought to be."



Inferring Norms from Numbers: Boomerang Effects of Online Virality Metrics on the Persuasiveness of HIV Prevention Messages

Christopher (CJ) Calabrese, MPH
Jingwen Zhang, PhD
Communication in Health and Technology Research Lab (CHATR) http://chatrlab.ucdavis.edu/
Department of Communication, University of California, Davis

UCDAVIS
UNIVERSITY OF CALIFORNIA

BACKGROUND

- Virality metrics depict the aggregated counts of online shares for a piece of information. Typically displayed as numbers alongside an article or a video, virality metrics indicate the real-time popularity of which information has attracted.
- A few studies experimentally identified its effects in influencing the persuasiveness of a health message through changing relevant behavior determinants including injunctive norm, perceived susceptibility, and fear. However, no research has investigated whether and how virality metrics and source similarity – another important online feature influencing message persuasiveness – interactively influence individuals' judgment of a health message and related behavioral determinants.
- To contribute to the literature, we examined how virality metrics impacted behavior intention through both injunctive norm and descriptive norm and whether source similarity moderated such effects in the context of a HIV prevention video message that promoted both condom use and HIV testing.

HYPOTHESES

- **H1. Virality metrics will lead to greater acceptance of the message than no virality metrics.**
- **H2. Virality metrics will lead to greater injunctive norm and descriptive norm supporting the recommended behavior than no virality metrics.**
- **H3. Virality metrics will lead to greater behavior intention through increasing injunctive norm and descriptive norm supporting the recommended behavior than no virality metrics.**
- **H4. Source similarity will moderate the effects of virality metrics, making the effects larger when the source matches the viewer's demographic characteristics including gender, age, and race.**

METHODS

- **Design**: An online factorial experiment testing the effects of virality metrics (i.e., high, medium, low, and no metric) and source similarity (i.e., similar versus dissimilar).
- **Sample**: 621 participants aged 18 to 35 recruited from Amazon mTurk in 2017.

ONLINE EXPERIMENT

The campaign video on HIV risk reduction

Manipulation of video poster

Manipulation of virality metrics

We could
all
reduce transmission
0:57 / 1:15
Transmission of HIV can be reduced by 96%
Arjun Patel shared this video
This video has been shared 1.5K times
1.5k
I would like to share this video:
Share
Tweet
G+1
Related Videos being reviewed by other participants:
THE END OF AIDS.
Living in the City
This video has been shared 32.8k times
32.8k
Safe Travels
This video has been shared 219 times
219
The Red Name
There is no information on sharing at this time

Note: These mediation models were analyzed using SPSS PROCESS Macro.
* p < .05, ** p < .01, *** p <. 001.

RESULTS

**
*
*
4
3
2
1
0
Condom Use Injunctive Norm
Condom Use Descriptive Norm
Condom Use Intention
HIV Testing Injunctive Norm
HIV Testing Descriptive Norm
HIV Testing Intention
No Virality Metrics
Virality Metrics

Note: * p < .05, ** p < .01.

1. Virality metrics DID NOT have an effect on message acceptance.
2. Virality metrics ***decreased participants' perceived injunctive norm for condom use*** (b = -.35, p = .004), and ***decreased intention to use condoms*** (b = -.33, p = .027).
3. Virality metrics ***decreased perceived injunctive norm for HIV testing*** (b = -.23, p = .031).
4. Source similarity moderated the effects of virality metrics on message acceptance (b = 1.07, p < .01). ***Virality metrics increased message acceptance only when the message source was matched with the user's profile.***

DISCUSSION

1. In contrast to what previous studies found, the strong boomerang effect shown in this study invites researchers to rethink about using virality metrics in online campaigns.
2. Virality metrics can backfire when people infer that others actively share the message because they are not encouraged by norms of protective behaviors and they want to make a change.
3. Virality metrics may enhance message acceptance only when people perceive the message poster share similar demographic characteristics with them.
4. These findings suggest a strong need for future replications studies. The boomerang effects may be replicated using similar persuasive messages addressing stigmatized health topics.

ACKNOWLEDGEMENT

We thank The Elizabeth Taylor AIDS Foundation for providing the campaign video for this study.

Confidence intervals tell us precisely how confident we can be about our sample proportion as it relates to our real population parameter. In the case of p-values, they only tell us whether or not a result from a sample is statistically significant, but not how significant it may be. It only shows us if our hypothesis test was statistically significant or, in other words, if we can reject the null hypothesis, which states that whatever we're testing had no significant impact on our sample. In the case of this research topic, our null hypothesis would be that virality metrics neither increase or decrease message effectiveness in online video campaigns. As section 3.11.4 in our textbook demonstrates, if we focus more on confidence intervals rather than p-values, we're able to measure the likelihood that our sample proportion is reasonably accurate by the width of the interval, while additionally being able to discern the significance of our hypothesis by observing the location of the interval itself. In the context of this study, confidence intervals might provide a more insightful analysis into the effects of virality metrics on both injunctive and descriptive norms, as well as to what extent we can be confident in these results. In the case of using p-values for this study, we run into the danger of skewing reasonably far from the true population proportion, and thus risk producing false positives in the case of multiple independent hypotheses (which are present in this experiment). This stems from the fact that we can't measure the degree of confidence of our sample proportion when only considering p-values. For example, this study's hypotheses regarding reduced condom use retention was shown to be false, accompanied by a p-value less than 0.05. This could have been derived from an inaccurate sample proportion, but it's much more difficult to tell without the use of confidence intervals.

2. Problem B

2.1. Data Wrangling:

All data manipulation/filtering was performed before splitting our data into `tstdta` and `trndta` to make sure the size of our holdout set was not affected.

Price Conversions:
To prepare the Airbnb listing data for our model and subsequent cross validation, we first converted the price column from strings representing monetary values into simply numbers. This was accomplished by replacing all dollar signs and commas with empty strings for every member of any relevant column which contained monetary values (most notably: price). Afterwards, we were able to convert this string to a number using as.numeric.

Invalid/Irrelevant Values:
As pointed out in the project description, certain price values may be erroneous or missing, thus we decided to exclude any rows with prices with a value of "NA" or those below the nightly minimum of $10 (according to Airbnb's website). We additionally excluded from our data the listing that had a country_code of "MX". Apparently, there is a city called San Francisco in Mexico, and we didn't want this data to skew our model.

String Data to Weighted Numerics
Since we believed that variables like room_type and bed_type would be semi-accurate predictors of price, we decided to make new variables corresponding to "weights" based on perceived value of each of these columns' options. For example, a "Real Bed" would be worth the most, and "Couch" would be worth the least in terms of sleeping accommodations. This way, we were able to include these new columns as quantitative assessments of formerly qualitative values.

Missing Values:
We grappled with how to handle missing values, as we weren't sure whether to set a default value to "NA" entries or ignore their rows entirely. We ultimately decided to exclude those rows, as including a default value worsened our fit. For example, we thought that number of bathrooms, security deposit, and cleaning fee were all important factors in deciding price, but many rows were found with "NA" entries for at least one of those 3 columns, and we didn't think it was completely sound to assume that a missing value indicated a 0 in those cases.

2.2. Predictor Choices:

All in all, we did not decide to utilize any polynomial terms because none of our data had negative values associated with them.

For our model, we used a combination of different variables within the Airbnb data that was provided. Majority of the variables we choose were reasoned from personal experience and prior knowledge of the real estate industry.

- bathrooms: Number of bathrooms impacts the square footage of the property available to the customer which influences pricing.
- bedrooms: Number of bedrooms impacts the square footage of the property available to the customer which influences pricing.
- room_type: Room type impacts the square footage of the property available to the customer which influences pricing. Since the data is only available as strings, we changed it to weighted numerals. The types of rooms available and weights used are as follows: entire home/apt - 3, private room - 2, shared room - 1. As can be deduced, higher weights were given to rooms that have more space.
- bed_type: Properties with a more comfortable bed type can price their listings higher as customers are willing to pay for comfort. Since the data is only available as strings, we changed it to weighted numerals. The types of beds available and weights used are as follows: real bed - 5, airbed - 4, futon - 3, pull-out sofa - 2, couch - 1. As can be deduced, higher weights were given to beds that have more generally comfortable.

- accommodates: Number of people accommodated impacts square footage of the property as more space is needed for more people. Thus, this variable affects pricing.

- cleaning_fee: Our inclusion of cleaning fee had multiple lines of reasoning. We thought it was likely that the higher the cleaning fee, the larger the property would be as it would be more expensive to clean larger spaces than smaller ones. Also, generally, cleaning fees are not larger than the property's price as it is simply unreasonable to do so. Based on this logic, we can say the higher the cleaning fee, the higher the price.

- security_deposit:

- review_scores_rating: Review ratings were included as we reasoned that properties with higher scores could afford to price their listing higher without deterring customers as customers tend to place a large emphasis on ratings when selecting their booking.

- review_scores_location: The higher the location rating, the more pleased customers are with the distance to attractions, conference buildings, and other places of interest. Thus, a listing can be priced higher if they have positive ratings in this category as people will still book with them for the location perk.

- review_scores_cleanliness: With a high cleanliness rating, listings can be priced higher as this is an aspect people generally value a great deal. When comparing a location that is cheap but unsanitary and a location that is more expensive but clean, base on personal experience, a customer is more likely to choose the more expensive but clean option.

- review_scores_communication

- review_scores_accuracy

- zipcode: We use zip code to determine the location of a property. Location matters as properties near water, in downtown, or in tourist hot spots tend to have higher prices. Though zip codes are numeric values themselves, we needed to weight them.....

Before settling on these parameters for our model for the testing data, we went through a number of iterations where we tried including different variables that are part of each listing. With that being said, we also attempted to use various variables that had character values and assign them to a scale of numbers. For certain variables, this still did not make a large enough difference for it to be included in the parameters we chose for our model.

- amenities: This was considered; We thought that the number of amenities provided would be important and possibly positively correlated to the price of the listing, but we were mistaken. We ran it against the test data, and the addition of this parameter hurt the mean absolute prediction error. If we were to use the amenities column, it would have likely been for individual features that may be more desirable, such as Wifi or a full kitchen.

- extra_people: Initially, we believed this variable would be relevant. But after consideration and testing with the test data, we realized that it was not relevant in determining the price of the listing.

- number_of_reviews: We originally used this value in our model; however, after doing a bit more research, we realized that more specific variables mattered more. After testing with our data and model, we determined that "number_of_reviews" should be replaced with "review_scores_cleanliness" as well as "review_scores_location". Cleanliness ratings and location ratings were far more relevant in determining a price for the listing. As a person searching for an Airbnb to book, we felt that cleanliness and location were two of the most important factors in making this decision.

2.3. Conclusion:

3. Group Member Contributions

1. Lynn Miyashita
   Wrote Problem A part 1. Wrote up the model choices section in this report.

2. Alisha Nanda
   Wrote Problem A part 2. Improved model quality through filtering missing values from our `lsts` data frame, wrote `weightBedTypes()` function to assist with quantitative analysis of room type values.

3. Sabrina Sheu
   Wrote Problem A part 3. Wrote `weightRoomTypes()` function to assist with quantitative analysis of room type values. Wrote the majority of the reasoning behind the chosen predictors.

4. Sam Waters
   Wrote Problem A part 4. Added initial data-importing and data-manipulation code, wrote the majority of the "Data Wrangling" section in the report, wrote `pricesToNumbers()` function to prepare our data frame's price column for cross validation. Code refactoring/cleanup – inclusion of switch statements, composing code sections into functions.

A. Code Listings

```r
this.dir <- dirname(parent.frame(2)$ofile)
setwd(this.dir)

airbnb <- function() {

  file <- "listings.csv"
  lsts <- read.csv(file, sep = ",",header = T, stringsAsFactors = FALSE)

  # Convert string monetary values to numbers
  lsts <- pricesToNumbers(lsts)

  # cat("length of lsts before filtering: ", length(lsts$price), "\n")

  invalidRows <- numeric(0)
  # Remove all rows with invalid values (prices less than $10, no reviews, certain NA
  for (i in 1:nrow(lsts)) {
    invalidRows <- findInvalidRows(i, lsts, invalidRows)
    lsts <- weightRoomTypes(i, lsts)
    lsts <- weightBedTypes(i, lsts)
  }

  lsts <- lsts[-invalidRows, ]

  lsts$accommodatesSq <- lsts$accommodates^2
  # print(lsts$bedrooms^2)
  # cat("length of lsts after filtering: ", length(lsts$price), "\n")

  set.seed(9999)
  idxs <- sample(1:nrow(lsts),1000)

  # Test data (holdout set)
  tstdta <- lsts[idxs,]
  # Training data
  trndta <- lsts[-idxs,]

  # Run lm on training set
  model <- lm(trndta$price ~ .,
              data = subset(trndta, select=c(cleaning_fee,
                                             security_deposit,
                                             bedrooms,
                                             bathrooms,
                                             review_scores_cleanliness,
                                             review_scores_accuracy,
                                             review_scores_communication,
                                             review_scores_location,
                                             review_scores_rating,
                                             accommodates,
                                             room_type_num,
                                             bed_type_num,
                                             accommodatesSq)),
              na.action = na.exclude)

  # Predict on the test set
  predvals <- predict(model, tstdta)

  #print(predvals)

  cat("compare␣against␣tstdta␣prices:␣", mean ( abs ( predvals - tstdta[ ,61])), "\n")

  trndtaSample <- sample(trndta$price,1000)
  cat("compare␣against␣actual␣trndta␣prices:␣", mean ( abs ( predvals - trndtaSample))
}

pricesToNumbers <- function(lsts) {
  lsts$price <- as.numeric(gsub('[$,]', '', lsts$price))
  lsts$extra_people <- as.numeric(gsub('[$,]', '', lsts$extra_people))
  lsts$security_deposit <- as.numeric(gsub('[$,]', '', lsts$security_deposit))
  lsts$cleaning_fee <- as.numeric(gsub('[$,]', '', lsts$cleaning_fee))
  return(lsts)
}

findInvalidRows <- function(i, lsts, invalidRows) {
  if (is.na(lsts$price[i])
      || is.na(lsts$bathrooms[i])
      || lsts$price[i] < 10
      || is.na(lsts$security_deposit[i])
      || is.na(lsts$cleaning_fee[i])
      || lsts$country[i] == "MX"
      || is.na(lsts$review_scores_rating[i])) {
    invalidRows <- c(invalidRows, i)
  }
  return(invalidRows)
}

weightBedTypes <- function(i, lsts) {
  lsts$bed_type_num[i] <- switch(lsts$bed_type[i], "Real␣Bed"=5, "Airbed"=4, "Futon"=3
  return(lsts)
}

weightRoomTypes <- function(i, lsts) {
  lsts$room_type_num[i] <- switch(lsts$room_type[i], "Entire␣home/apt"=3, "Private␣roo
  return(lsts)
}
```

References

[1] Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). *A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places.* Journal of Personality and Social Psychology, 58, 1015

- 1026. doi:10.1037/0022-3514.58.6.1015

[2] Convert currency with commas into numeric
https://stackoverflow.com/questions/31944103/convert-currency-with-commas-into-numeric

[3] R Switch Statement
https://www.tutorialgateway.org/r-switch-statement/m