

Predicting Loneliness from Social, Technology, and Demographic Factors

Jean Singer

9/4/2022

Contents

1. Introduction	2
1.1 Background	2
1.2 Project Objectives	2
1.3 Dataset and Variables	2
1.4 Key Steps Performed	3
2. Methods and Analysis	3
2.1 Exploratory Analysis and Variable Selection	3
2.1.1 Exploratory Analysis of Outcome Variable	5
2.1.2 Exploratory Analysis of Predictors and Variable Selection	8
2.2 Modeling Approaches	17
3. Models and Results	18
3.1 Linear Regression Model	19
3.2 XGBoost Model	21
3.3 Principal Components Regression Model (PCR)	23
3.4 Ensemble Model	25
3.5 Variable Importance	26
4. Conclusion	27
4.1 Discussion of Results and Potential Impact	27
4.2 Limitations	28
4.3 Future Work	28
Appendix A: AARP Variables	29
References	30

1. Introduction

1.1 Background

Loneliness in the US has reached what has been described as epidemic proportions. In a 2020 study of more than 10,400 adults by the health insurer Cigna, roughly three in five Americans (61%) were classified as lonely (Coombs 2020). Compared to 2018, when the survey was first introduced, loneliness rates rose by roughly 7 percentage points. Studies using different methodologies show various rates of loneliness but confirm that a sizeable proportion of the country suffers from this condition. The Covid-19 pandemic may also be playing a role in the rise of loneliness. In October 2020, Harvard University researchers found that 36% of Americans felt “serious loneliness” while two months prior, the proportion was 25% (Weissbourd et al., n.d.).

Social connection is a fundamental human need, akin to our needs for food and warmth (Lieberman 2013). When our social needs are insufficiently met, we feel loneliness, which has been defined as “a distressing feeling that accompanies the perception that one’s social needs are not being met by the quantity or especially the quality of one’s social relationships” (L. Hawkey and Cacioppo 2010). Loneliness is a signal that we need to improve social connection—and our biology responds if we don’t. Chronic loneliness has been associated with increased risk for a host of physical and mental health issues, including a 26% increase in the risk for early mortality, a rate comparable to those for obesity, inactivity, and smoking 15 cigarettes a day (Holt-Lunstad, Smith, and Layton 2010a; Holt-Lunstad et al. 2015). Loneliness has also been associated with increased risks for coronary heart disease and stroke (Everson-Rose and Lewis 2005), impaired immune function (Cohen et al. 1997a), cognitive decline and dementia (Cacioppo and Cacioppo 2014; Kim 2017). Given the reported rates of loneliness and the prospect that it may be further on the rise, we need to address loneliness as a serious public health concern.

Loneliness is a *measurable feeling* that may be influenced by a number of *objective factors*, such as the size or diversity of our social networks, the amount of time we spend alone, or the ways in which we use technology. If we can better understand the drivers behind loneliness and quantify their impact, we can more effectively target the factors that matter most. Machine learning models that predict loneliness from social, technology and demographic factors can help us to identify relevant factors to target.

1.2 Project Objectives

The objective of this project was to build a machine learning model to predict loneliness from a set of social, technology and demographic factors. The data I used was collected by AARP for a study of loneliness among people age 45 and above. AARP performed various analyses on their data, including a linear regression model predicting loneliness.

In the current project, I sought to:

- Build a linear regression model with better predictive ability than the one built by AARP.
- Further optimize predictive ability with three additional machine learning models: XGBoost, PCR and an ensemble.
- Use machine learning models to identify the social, technology and demographic features most important to predicting loneliness.

1.3 Dataset and Variables

In 2018, AARP conducted a study of loneliness and how it relates to social connections, life experiences, health, and technology among adults age 45 and over. They contacted a nationally-representative sample of 6343 people for participation in a web-based survey, and obtained a response rate of 50.8%, or 3223 people. I

removed from the dataset 20 people (0.6%) who did not complete the key outcome measure (UCLA Loneliness Index.) The final dataset consisted of 3203 observations.

The main outcome variable was loneliness, measured by the UCLA Loneliness Scale (Russell 1996), a validated and widely-used index of 20 items measuring dimensions of loneliness on a 4-point scale from “never” to “always.” The index is created by summing the responses. The possible range of responses is 20 (least lonely) to 80 (most lonely), and AARP classifies a score of 44 or higher as “lonely.”

The AARP dataset I started with contained 67 predictor variables. The variables included measures of social connection, technology use, life events such as moving or death of someone close, activities such as volunteering and belonging to groups, attitudes regarding the internet, health, and demographics. A list of the variables is provided in Appendix A: AARP Variables.

The AARP variables did not include a measure of complex social integration (CSI), which is an indicator of the extent to which an individual participates in a wide range of social activities and relationships (Brisette, Cohen, and Seeman 2000). CSI has been frequently associated with increased risk of mortality (Holt-Lunstad, Smith, and Layton 2010b; L. F. Berkman and Syme 1979a; Lisa F. Berkman et al. 2004) and morbidity (Cohen et al. 1997b; Coyle and Dugan 2012), and is a commonly-used measure in studies of social connection and isolation. I created an index of CSI using the variable DiversitySupportive (a measure of the diversity of people who have been supportive of you in the last year, across friends, spouse, children, parents, other relatives, neighbors, co-workers, and others) and adding points for being married or living with a partner, being employed, volunteering, membership in groups. This approach is similar to those used by researchers to create other established indexes such as the Berkman-Syme Social Network Index (L. F. Berkman and Syme 1979b) and the Cohen Social Network Index (Cohen et al. 1997c).

Here I load the dataset and add the CSI variable. (Code not shown in pdf version.)

1.4 Key Steps Performed

I started with a process of exploring the data and selecting the variables that were most likely to be predictive in my models. Using those variables, I built three machine learning models: linear regression, XGBoost (Extreme Gradient Boosting), and principal components regression (PCR.) I evaluated my linear regression model against AARP’s linear regression model using R^2 as my performance measure. I also evaluated each of my models against each other using RMSE. (AARP did not report RMSE for their model.) I combined the three models into an ensemble and compared RMSE to the individual models.

I used the linear regression and XGBoost models to generate measures of variable importance. (PCR and the ensemble do not provide variable importance metrics.) I then examined the top five most important variables and compared them between the two models to see were they found the same variables to be important.

2. Methods and Analysis

2.1 Exploratory Analysis and Variable Selection

I examined the structure of the data. The dataset contains 3203 rows and 70 variables: 68 predictor variables, one outcome variable and the respondent number. All variables are shown as numeric, although some should be factor type. I will convert them later on.

```
str(df)
```

```
## 'data.frame':    3203 obs. of  69 variables:
## $ respondent      : int   3 4 5 6 7 8 9 10 11 12 ...
## $ UCLA_index      : num   29 23 51 31 48 28 36 32 37 39 ...
## $ Q8_disability    : int    0 0 0 0 1 0 0 1 0 1 ...
```

```

## $ NeighborIndex : num 0.0871 2.2295 2.0115 -0.3488 2.2295 ...
## $ DiversityDiscussImportant : int 3 2 1 5 NA 4 4 1 3 3 ...
## $ DiversitySupportive : num 5 4 1 6 NA 3 2 1 2 4 ...
## $ Q22_marital_satn : int 5 5 5 5 3 5 2 5 1 0 ...
## $ Q46_attend_religious : int 6 6 6 2 1 1 3 2 2 3 ...
## $ Q39_caregiver : int 0 0 1 0 0 0 0 0 0 0 ...
## $ Q30_supportive_num : int 100 12 5 8 0 5 2 20 2 25 ...
## $ Q64_yrs_current_residence : int 4 1 5 2 1 2 5 2 5 3 ...
## $ Q65_relocate : int 0 1 0 3 1 2 0 1 0 1 ...
## $ ethnic : int 1 1 1 1 1 1 1 1 1 1 ...
## $ gender : int 2 1 1 1 1 2 1 1 1 2 ...
## $ household_size : int 4 2 1 4 3 1 2 2 2 1 ...
## $ income : int 14 19 21 9 16 9 21 17 21 6 ...
## $ employ : int 7 5 1 1 5 5 2 2 5 5 ...
## $ age_group : int 5 7 7 5 6 7 7 7 7 7 ...
## $ Q2_health_overall : int 5 4 5 2 2 5 3 2 4 4 ...
## $ Q27_11_parents_inperson : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_12_parents_email : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_13_parents_phone : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_14_parents_letters : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_15_parents_text : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_16_parents_online : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_17_parents_SN : int NA NA NA NA NA NA NA NA NA NA ...
## $ Q27_21_child_inperson : int 5 3 5 5 5 4 5 5 4 NA ...
## $ Q27_22_child_email : int 1 4 5 4 1 1 5 5 5 NA ...
## $ Q27_23_child_phone : int 5 4 5 5 5 5 4 5 5 NA ...
## $ Q27_24_child_letters : int 1 1 2 1 1 1 2 3 2 NA ...
## $ Q27_25_child_text : int 5 5 5 5 5 5 5 5 5 NA ...
## $ Q27_26_child_online : int 1 5 5 4 1 1 1 1 5 NA ...
## $ Q27_27_child_SN : int 1 4 5 5 1 5 4 1 4 NA ...
## $ Q27_31_sibling_inperson : int 5 2 4 4 2 4 4 3 4 4 ...
## $ Q27_32_sibling_email : int 1 3 4 3 1 1 4 4 5 5 ...
## $ Q27_33_sibling_phone : int 5 3 4 3 2 5 3 4 3 4 ...
## $ Q27_34_sibling_letters : int 1 1 2 1 1 1 2 1 2 4 ...
## $ Q27_35_sibling_text : int 5 4 4 4 1 5 2 4 5 5 ...
## $ Q27_36_sibling_online : int 1 4 4 1 1 1 1 1 3 1 ...
## $ Q27_37_sibling_SN : int 1 1 4 5 1 4 3 1 5 1 ...
## $ Q27_41_friend_inperson : int 4 5 5 5 3 4 5 5 4 5 ...
## $ Q27_42_friend_email : int 1 1 5 5 1 2 5 5 5 5 ...
## $ Q27_43_friend_phone : int 4 1 5 5 3 5 2 4 5 5 ...
## $ Q27_44_friend_letters : int 2 1 5 1 1 1 2 1 2 4 ...
## $ Q27_45_friend_text : int 5 5 5 5 1 4 2 4 5 4 ...
## $ Q27_46_friend_online : int 1 5 5 1 1 1 1 1 2 2 ...
## $ Q27_47_friend_SN : int 1 1 1 2 5 2 5 5 1 5 ...
## $ Q28_discuss_important_matters_num : int 3 2 1 5 0 4 6 1 2 5 ...
## $ passaway_index : int 1 0 1 1 NA 0 0 2 0 2 ...
## $ moveaway_index : int 0 0 1 0 0 1 0 NA 1 1 ...
## $ Q38_more_less_friends : int 2 3 2 3 2 2 3 2 2 2 ...
## $ Q48_volunteer : int 1 1 0 0 0 0 1 0 1 1 ...
## $ Q50_groups : int 1 3 1 1 1 1 1 2 2 3 ...
## $ Q53_hobbies : int 5 2 3 2 4 2 2 2 5 5 ...
## $ Q77_hrs_alone : int 2 1 1 1 3 1 3 1 1 4 ...
## $ Q89_1_internet_sentiment : num 1 2 5 4 1 5 4 2 4 3 ...
## $ Q89_2_internet_sentiment : num 1 2 5 4 1 4 3 3 1 1 ...

```

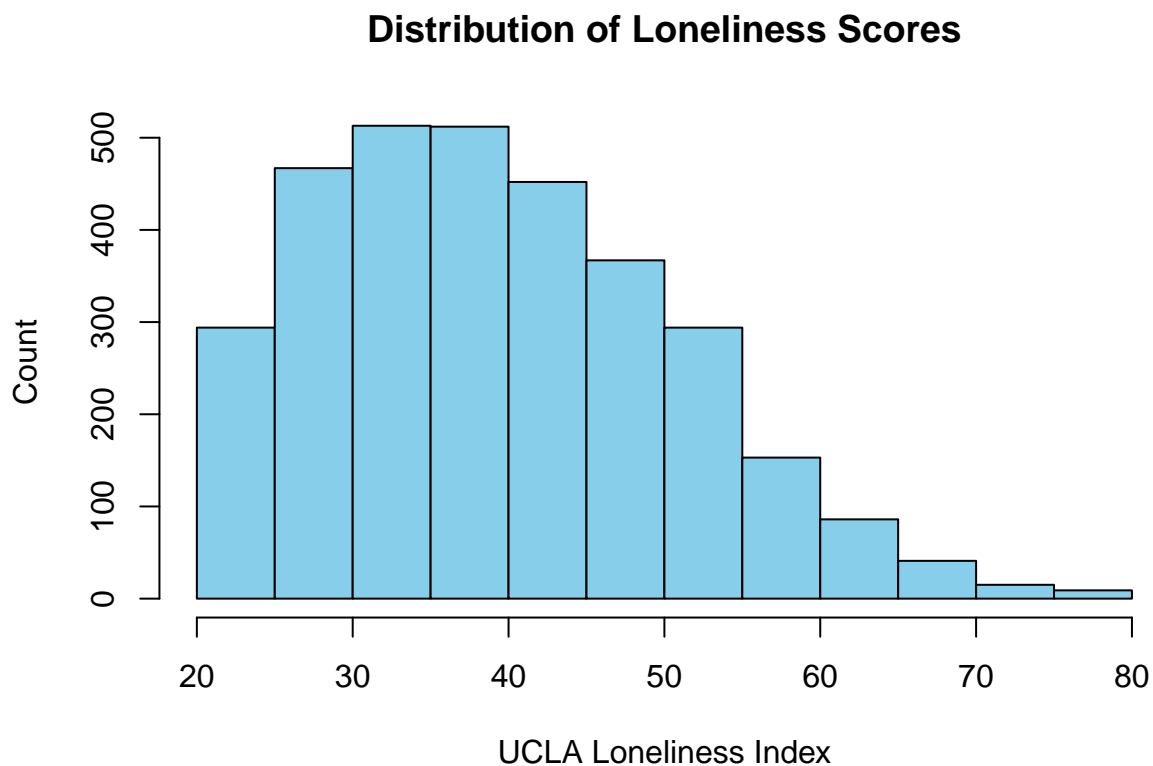
```
## $ Q89_3_internet_sentiment      : num  4 1 1 3 5 1 2 2 2 1 ...
## $ Q89_4_internet_sentiment      : int   5 4 5 5 1 5 4 1 3 2 ...
## $ Q89_5_internet_sentiment      : int   5 5 3 3 5 5 4 5 3 3 ...
## $ Q89_6_internet_sentiment      : num   5 4 3 5 1 5 3 1 3 3 ...
## $ Q89_7_internet_sentiment      : num   4 5 3 3 5 4 3 4 3 5 ...
## $ Q89_8_internet_sentiment      : num   5 1 3 5 1 3 3 2 3 3 ...
## $ Q89_9_internet_sentiment      : num   4 5 3 2 5 4 4 3 4 4 ...
## $ Q89_10_internet_sentiment     : num   5 5 3 4 3 3 5 2 1 3 ...
## $ internet_sentiment_index      : num  39 34 34 38 28 39 35 25 27 28 ...
## $ Q90_tradeoffs_family          : int    2 2 1 1 2 2 2 2 2 2 ...
## $ Q91_tradeoffs_intimate_convo  : int    2 2 1 1 2 2 2 2 2 1 ...
## $ CSI                           : num    7 7 3 8 NA 4 4 4 4 6 ...
```

2.1.1 Exploratory Analysis of Outcome Variable

For the main outcome variable—loneliness as measured by the UCLA Loneliness Index—I generated a histogram, examined the distribution, and examined incidence of loneliness by age.

Here is the histogram:

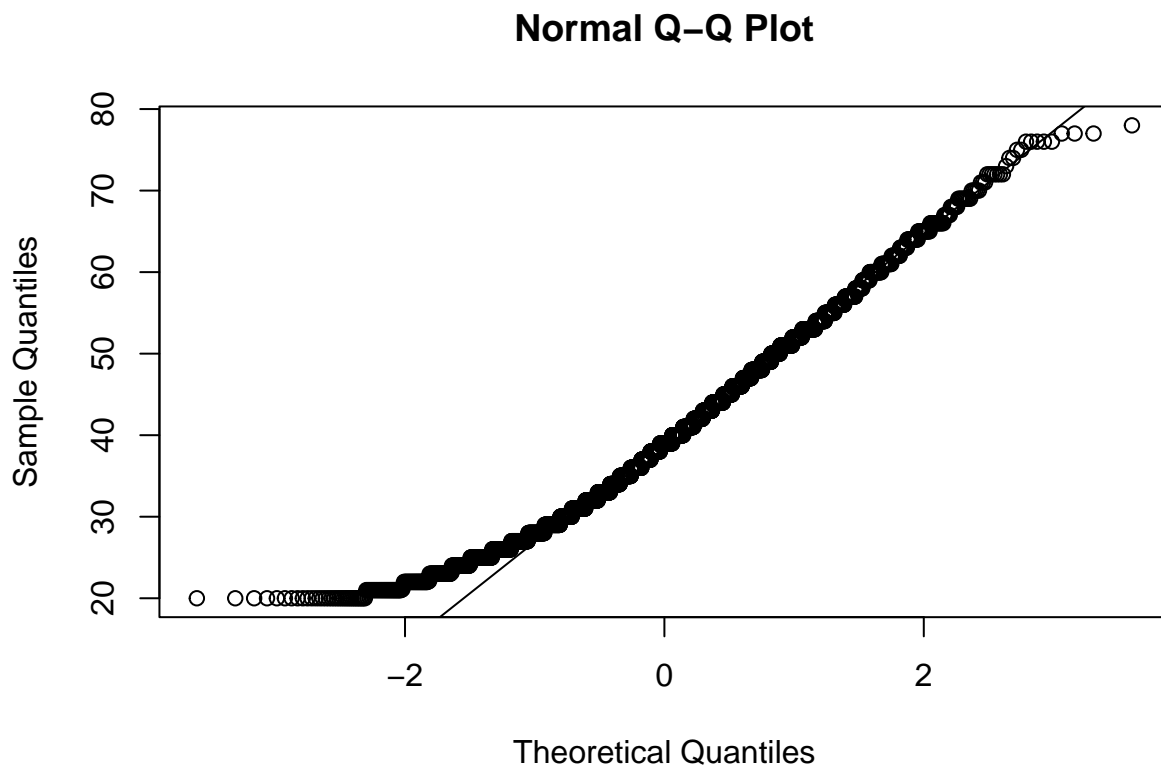
```
#create a histogram of the loneliness variable
hist(df$UCLA_index, main = "Distribution of Loneliness Scores",
      xlab = "UCLA Loneliness Index", ylab = "Count", col = "sky blue", border = "black")
```



The loneliness histogram is skewed right. A relatively large number of respondents are at the lowest end of the scale compared to the number at the highest end of the scale.

This deviation from the normal distribution can also be seen in the Q-Q plot. The distribution follows the Q-Q line fairly well except for the lowest values.

```
#create a Q-Q plot of loneliness
qqnorm(df$UCLA_index)
qqline(df$UCLA_index)
```



The chart below shows the mean, standard deviation, median and mode values for the loneliness variable. The mean is higher than the median, as would be expected in a right-skewed distribution.

```
#compute the mean, sd, median and mode for loneliness
mean_loneliness <- mean(df$UCLA_index)
sd_loneliness <- sd(df$UCLA_index)
median_loneliess <- median(df$UCLA_index)
library(DescTools)
mode_loneliness <- Mode(df$UCLA_index)

#place the stats in a table
descriptive_stats <- tibble(Statistic = c("Mean", "SD", "Median", "Mode"),
                                Value = c(mean_loneliness, sd_loneliness,
                                           median_loneliess, mode_loneliness))
descriptive_stats %>% knitr::kable()
```

Statistic	Value
Mean	39.80182
SD	11.34231
Median	39.00000
Mode	39.00000

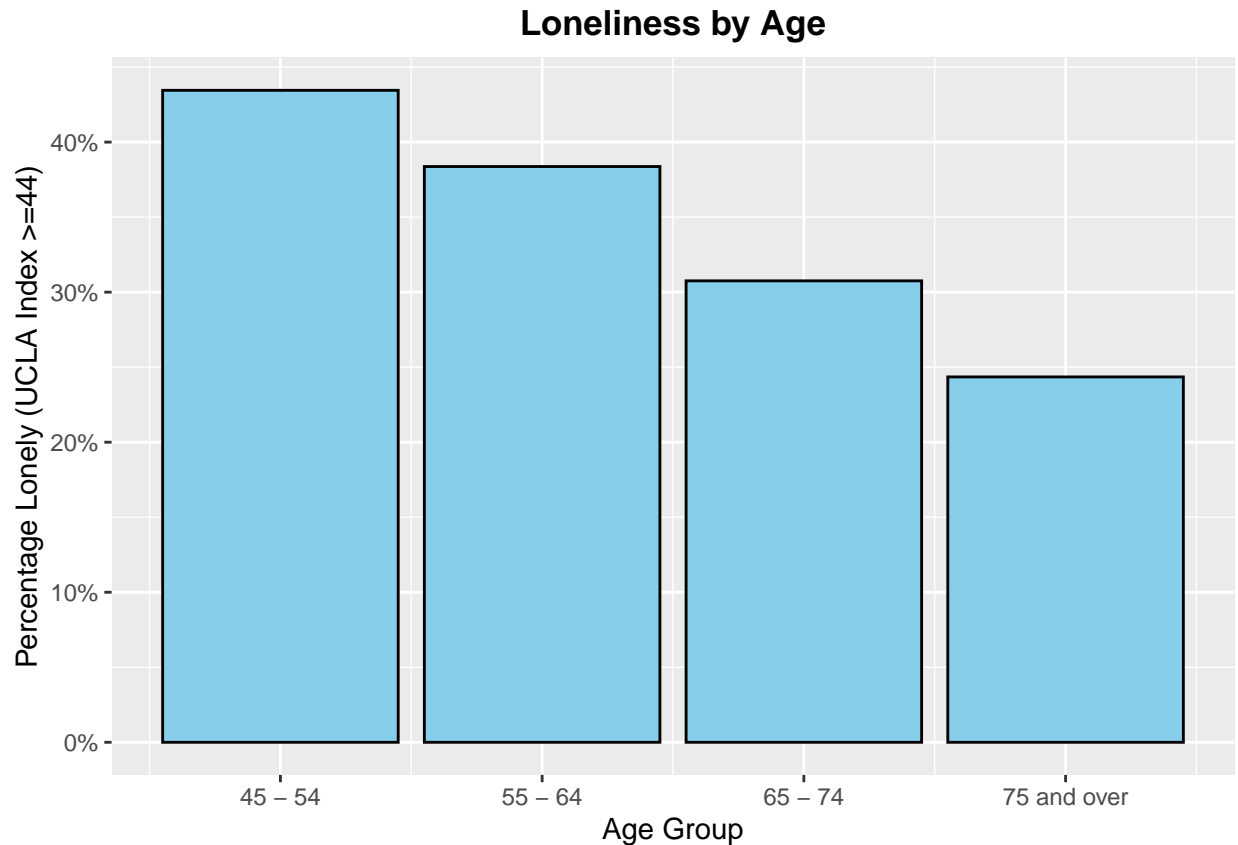
Using the AARP cutoff of 44 and above to define “lonely”, I found that a little over a third (35.6%) of the sample would be considered lonely.

```
#compute the percentage of people who would be considered lonely (UCLA index >=44)
mean(df$UCLA_index >=44)
```

```
## [1] 0.3562285
```

A plot of average loneliness by age group shows that loneliness declines with age. This finding is consistent with other studies indicating that loneliness is highest among young people (around age 16 – 24) and declines after middle age, although some studies show loneliness increasing at age 75 and above (L. C. Hawkey et al. 2019; “Community Life Survey 2018-19,” n.d.).

```
#Percentage of people classified as "lonely" by age
df %>% group_by(age_group) %>%
  summarise(Percentage_lonely = mean(UCLA_index >= 44)) %>%
  ggplot(aes(age_group, Percentage_lonely)) + geom_bar(stat = "identity",
                                                    col = "black", fill = "sky blue") +
  xlab("Age Group") + ylab("Percentage Lonely (UCLA Index >=44)") +
  labs(title = "Loneliness by Age") + theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  scale_y_continuous(labels = scales::percent_format(accuracy=1)) +
  scale_x_continuous(breaks = 5:8, labels = c("45 - 54", "55 - 64", "65 - 74", "75 and over"))
```



2.1.2 Exploratory Analysis of Predictors and Variable Selection

To identify the predictor variables that would be most useful in predicting loneliness, I conducted three types of analyses: correlation with loneliness, ANOVA, and t-tests.

Correlation with loneliness. I treated ordinal variables with 5 or more response options and any index variables as “pseudo-continuous”, and generated Pearson’s correlation coefficients for them. I selected all variables with correlations of greater than .20 or less than -.20, and $p < .05$ for use in my machine learning models. Out of 55 variables, only 13 met this fairly low hurdle, indicating that individually, the variables were not very strong predictors. All p-values, however, were significant ($p < .001$). The variables that met these criteria are shown in the table below.

Below is the code for running the correlations and selecting the variables that meet my criteria.

```
#Create a correlation matrix for the variables of interest, which are
#those that are indexes or ordinal with 5 or more response options
variables <- df %>% select(-c(respondent, ethnic,
                             Q22_marital_satn, Q38_more_less_friends,
                             Q50_groups,
                             Q90_tradeoffs_family, Q91_tradeoffs_intimate_convo,
                             age_group, employ, gender,
                             Q39_caregiver, Q48_volunteer, Q8_disability))

#Generate the correlation matrix and p values
library(Hmisc)
cor_variables <- rcorr(as.matrix(variables))
```



```

cor_variables_r <- round(cor_variables[["r"]],2)
cor_variables_p <- round(cor_variables[["p"]],3)

#In the r-value matrix, select the variables with correlations >.20 or <-.2
new_variables <- as.data.frame(cor_variables_r) %>%
  filter(UCLA_index>0.2 | UCLA_index<(-.2)) %>%
  select(UCLA_index)

#create a column with the names of all the variables
new_variables$variable_names <- row.names(new_variables)

#remove the first row that contains UCLA_index as a variable
new_variables <- new_variables[-1, ]

#Now take a look at the p-values
#In the p-value matrix, add variable names as a column
new_variables_p <- as.data.frame(cor_variables_p)
new_variables_p$variable_names <- row.names(new_variables_p)

#select the relevant variables from the p-value matrix
relevant_variables <- c("NeighborIndex", "DiversityDiscussImportant",
  "DiversitySupportive", "Q30_supportive_num",
  "income", "Q2_health_overall",
  "Q27_41_friend_inperson",
  "Q27_43_friend_phone",
  "Q28_discuss_important_matters_num",
  "Q77_hrs_alone", "Q89_5_internet_sentiment",
  "Q89_7_internet_sentiment", "CSI")

new_variables_p <- new_variables_p %>% filter(variable_names %in% relevant_variables) %>%
  select(UCLA_index, variable_names)

#add three decimal places to the p-value
new_variables_p[, 'UCLA_index'] = format(round(new_variables_p[, 'UCLA_index'], 3), nsmall=3)

#create a table with the variables, correlations and p-values for all variables
# with r>0.20
correlation_table <- tibble(Variable_name = new_variables$variable_names,
  Pearsons_r = new_variables$UCLA_index,
  p_value = new_variables_p$UCLA_index)

#sort the table
correlation_table <- correlation_table %>% arrange(desc(Pearsons_r))

#format the table
correlation_table %>% knitr::kable()

```

Variable_name	Pearsons_r	p_value
Q77_hrs_alone	0.33	0.000
Q27_43_friend_phone	-0.21	0.000
income	-0.22	0.000
Q30_supportive_num	-0.23	0.000
DiversityDiscussImportant	-0.24	0.000

Variable_name	Pearsons_r	p_value
Q89_7_internet_sentiment	-0.24	0.000
Q89_5_internet_sentiment	-0.26	0.000
Q28_discuss_important_matters_num	-0.27	0.000
Q2_health_overall	-0.28	0.000
Q27_41_friend_inperson	-0.29	0.000
NeighborIndex	-0.32	0.000
DiversitySupportive	-0.33	0.000
CSI	-0.37	0.000

We can see from the table above that the only predictor positively associated with loneliness is number of hours spent alone (Q77_hrs_alone.) The positive relationship implies that the more hours we spend physically alone, the lonelier we feel. The strongest predictor overall (although still not a “strong” correlation) is complex social integration (CSI), which is negatively associated with loneliness. The negative relationship implies that the more we participate in a wide range of social activities and relationships, the less lonely we are.

I generated scatterplots of the 13 variables but because the correlations were low, the scatterplots were not informative. (Plots not shown.)

ANOVA. For variables that were nominal or that were ordinal with fewer than 5 responses, I ran ANOVAs to determine whether there was a statistically significant difference ($p < .05$) in mean loneliness across each of the categorical groups. The 7 relevant variables were:

- ethnic (ethnicity)
- Q22_marital_satn (marital satisfaction)
- Q38_more_less_friends (increase or decrease in number of friends over the past five years)
- Q50_groups (number of groups you belong to)
- Q90_tradeoffs_family (whether you spend more/less/same time with family as a result of the internet)
- Q91_tradeoffs_intimate_convo (whether you spend more/less/same time in intimate conversations as a result of the internet)
- age_group.

All seven variables showed significant differences in mean loneliness among their categorical groups ($p < .01$ or better.) As a result, I included all of them in my models. Below is the code for running the seven ANOVAs.

```
#Convert all the variables to factor
df$Q22_marital_satn <- as.factor(df$Q22_marital_satn)
df$age_group <- as.factor(df$age_group)
df$Q38_more_less_friends <- as.factor(df$Q38_more_less_friends)
df$Q90_tradeoffs_family <- as.factor(df$Q90_tradeoffs_family)
df$Q91_tradeoffs_intimate_convo <- as.factor(df$Q91_tradeoffs_intimate_convo)
df$Q50_groups <- as.factor(df$Q50_groups)
df$ethnic <- as.factor(df$ethnic)

#ANOVA of Q22_marital_satn vs loneliness
df_marital_satn <- df %>%
  filter(Q22_marital_satn != "NA" & UCLA_index != "NA")
ANOVA_marital_satn_loneliness <- aov(UCLA_index ~ Q22_marital_satn,
                                     data = df_marital_satn)
summary(ANOVA_marital_satn_loneliness)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Q22_marital_satn    5  43169    8634   74.85 <2e-16 ***
## Residuals          3197 368762    115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#ANOVA of ethnicity vs loneliness

```
df_ethnic <- df %>%
  filter(ethnic != "NA" & UCLA_index != "NA")
ANOVA_ethnic <- aov(UCLA_index ~ ethnic,
  data = df_ethnic)
summary(ANOVA_ethnic)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ethnic          4   1766    441.6   3.443 0.00815 **
## Residuals       3198 410165    128.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#ANOVA of age group vs loneliness

```
df_agegroup <- df %>%
  filter(age_group != "NA" & UCLA_index != "NA")
ANOVA_agegroup_loneliness <- aov(UCLA_index ~ age_group,
  data = df_agegroup)
summary(ANOVA_agegroup_loneliness)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_group       3  11562    3854   30.79 <2e-16 ***
## Residuals       3199 400369    125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#ANOVA of Q38_more_less_friends (changes in number of friends #over the last 5 years) vs loneliness

```
df_friends <- df %>%
  filter(Q38_more_less_friends != "NA" & UCLA_index != "NA")
ANOVA_friends_loneliness <- aov(UCLA_index ~ age_group,
  data = df_friends)
summary(ANOVA_friends_loneliness)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_group       3  11561    3854   30.8 <2e-16 ***
## Residuals       3192 399309    125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#ANOVA of Q90_tradeoffs_family (as a result of technology, spending #less, same or more time with family) vs loneliness

```
df_tradeoffsfam <- df %>%
  filter(Q90_tradeoffs_family != "NA" & UCLA_index != "NA")
ANOVA_tradeoffsfam_loneliness <- aov(UCLA_index ~ Q90_tradeoffs_family,
  data = df_tradeoffsfam)
summary(ANOVA_tradeoffsfam_loneliness)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Q90_tradeoffs_family    2  18026    9013   73.19 <2e-16 ***
## Residuals              3185 392214    123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#ANOVA of Q91_tradeoffs_intimate_convo (as a result of technology, spending
#less, same or more time in intimate conversations) vs loneliness
df_tradeoffsint <- df %>%
  filter(Q91_tradeoffs_intimate_convo != "NA" & UCLA_index != "NA")
ANOVA_tradeoffsint_loneliness <- aov(UCLA_index ~ Q91_tradeoffs_intimate_convo,
                                     data = df_tradeoffsint)
summary(ANOVA_tradeoffsint_loneliness)
```

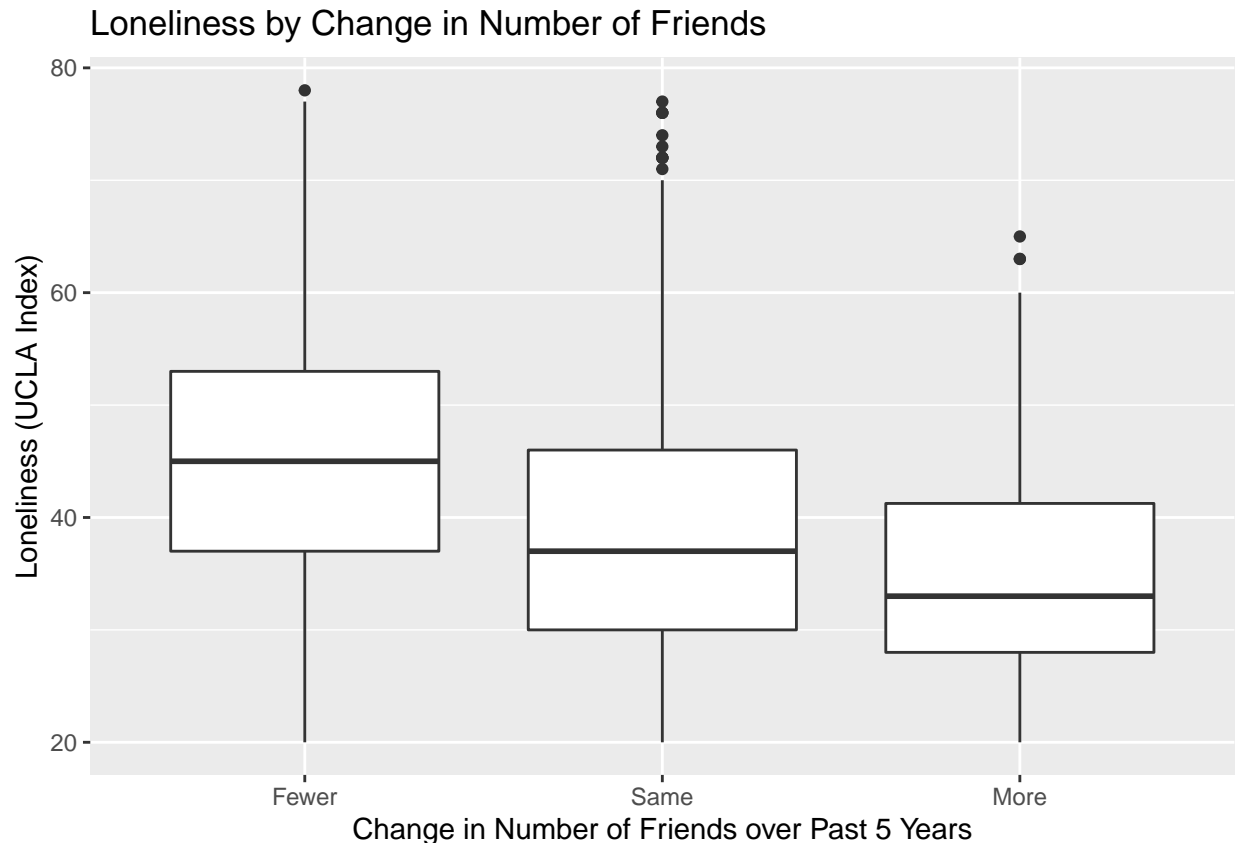
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Q91_tradeoffs_intimate_convo    2  19442    9721   79.61 <2e-16 ***
## Residuals              3119 380876    122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#ANOVA of Q50_groups (number of groups you belong to with
#1=0, 2=1, 3=2, 4=3 or more) vs loneliness
df_groups <- df %>%
  filter(Q50_groups != "NA" & UCLA_index != "NA")
ANOVA_groups_loneliness <- aov(UCLA_index ~ Q50_groups,
                              data = df_groups)
summary(ANOVA_groups_loneliness)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Q50_groups      3  11117    3706   29.52 <2e-16 ***
## Residuals      3187 400136    126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I generated boxplots for all seven variables. Below I show an example illustrating the association between loneliness and change in the number of friends over the last five years. The boxplot illustrates that loneliness on average is highest when the number of friends has declined, and lowest when the number of friends has increased. (The other six boxplots are not shown.)

```
df %>% filter(Q38_more_less_friends != "NA") %>%
  ggplot(aes(as.factor(Q38_more_less_friends), UCLA_index)) + geom_boxplot() +
  xlab("Change in Number of Friends over Past 5 Years") + ylab("Loneliness (UCLA Index)") +
  labs(title = "Loneliness by Change in Number of Friends")+
  scale_x_discrete(breaks = c("1", "2", "3"),
                  labels = c("Fewer", "Same", "More"))
```



T-TESTS. For binary variables, I conducted t-tests and selected those variables whose levels show significant differences ($p < .05$) in mean level of loneliness. Three variables showed significant differences and I included them in the models:

- Q48_volunteer (whether you have volunteered in the last 12 months), $p < .001$
- Q8_disability (whether you have a disability or chronic disease that keeps you from participating fully in work, school, household, or other activities), $p < .001$
- Q39_caregiver (whether you are providing unpaid care or assistance to an adult who needs assistance due to aging, a disability, or a health-related issue), $p < .01$

Two variables did not show significant differences between levels and were removed. These were gender ($p = 0.17$) and employed ($p = .06$).

Below is the code for running the five t-tests.

```
#t-test on mean loneliness for Q48_volunteer (whether or not you have
#in the last 12 months.)
df_volunteer <- df %>%
  filter(Q48_volunteer != "NA" & UCLA_index != "NA")
t.test(UCLA_index ~ Q48_volunteer, data = df_volunteer)
```

```
##
## Welch Two Sample t-test
##
```

```
## data: UCLA_index by Q48_volunteer
## t = 8.4587, df = 3029.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## 2.564041 4.111443
## sample estimates:
## mean in group 0 mean in group 1
## 41.17371 37.83597
```

```
#t-test on mean loneliness for Q8_disabled (whether or not you are disabled)
df_disabled <- df %>%
  filter(Q8_disability != "NA" & UCLA_index != "NA")
t.test(UCLA_index ~ Q8_disability, data = df_disabled)
```

```
##
## Welch Two Sample t-test
##
## data: UCLA_index by Q8_disability
## t = -9.2856, df = 932.58, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -6.00009 -3.90637
## sample estimates:
## mean in group 0 mean in group 1
## 38.77344 43.72667
```

```
#t-test on mean loneliness for Q39_caregiver (whether or not you are
#providing unpaid care to an adult)
df_caregiver <- df %>%
  filter(Q39_caregiver != "NA" & UCLA_index != "NA")
t.test(UCLA_index ~ Q39_caregiver, data = df_caregiver)
```

```
##
## Welch Two Sample t-test
##
## data: UCLA_index by Q39_caregiver
## t = -3.2898, df = 623.35, p-value = 0.001059
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -2.9645117 -0.7482516
## sample estimates:
## mean in group 0 mean in group 1
## 39.52254 41.37892
```

```
#t-test on mean loneliness for gender
df_gender <- df %>%
  filter(gender != "NA" & UCLA_index != "NA")
t.test(UCLA_index ~ gender, data = df_gender)
```

```
##
## Welch Two Sample t-test
##
```

```
## data: UCLA_index by gender
## t = 1.3622, df = 3194.3, p-value = 0.1732
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
## -0.2398954 1.3317958
## sample estimates:
## mean in group 1 mean in group 2
## 40.07522 39.52927
```

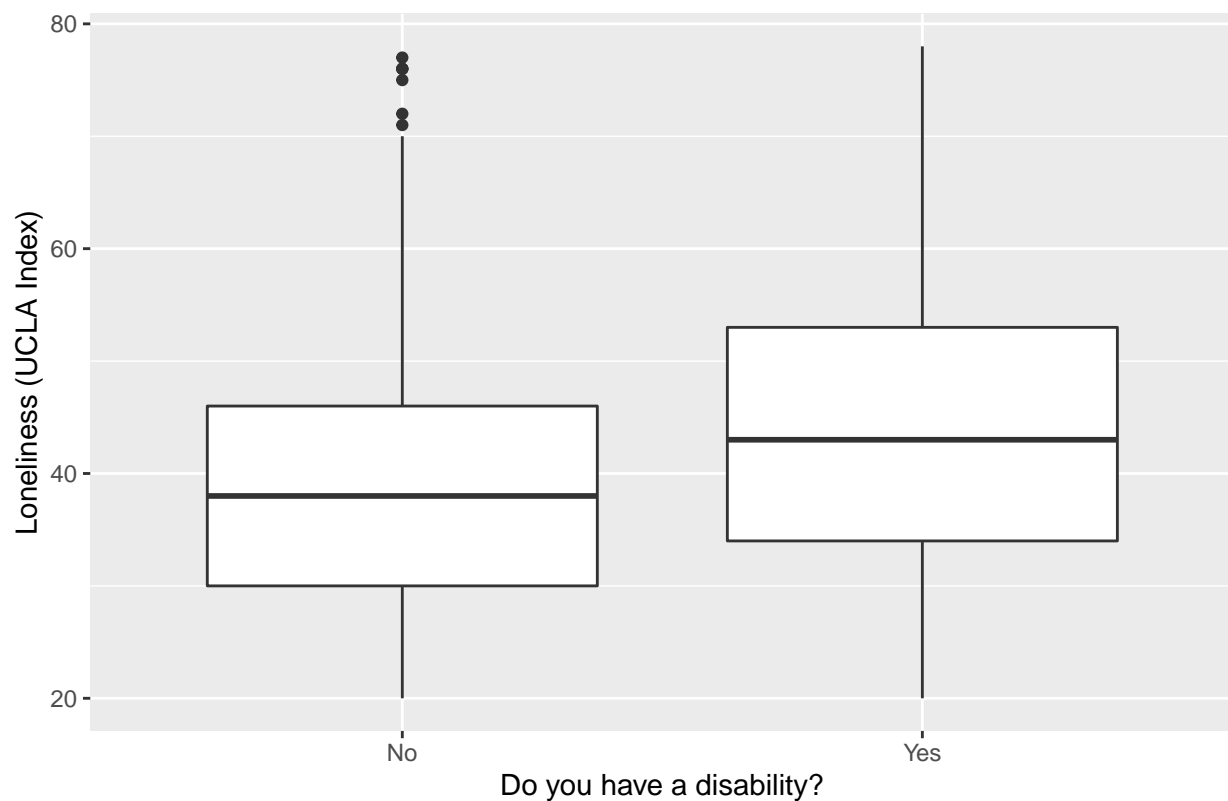
```
#t-test on mean loneliness for whether or not you're employed
df_employed <- df %>% mutate(CSIemployed = ifelse(employ==1 | employ==2, 1, 0))%>%
  filter(CSIemployed != "NA" & UCLA_index != "NA")
t.test(UCLA_index ~ CSIemployed, data = df_employed)
```

```
##
## Welch Two Sample t-test
##
## data: UCLA_index by CSIemployed
## t = -1.9196, df = 3133.6, p-value = 0.05499
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -1.55883759 0.01650072
## sample estimates:
## mean in group 0 mean in group 1
## 39.40697 40.17813
```

I generated boxplots on all five variables. Below I show examples for Q8_disability, where the means are significantly different, and gender, where they are not.

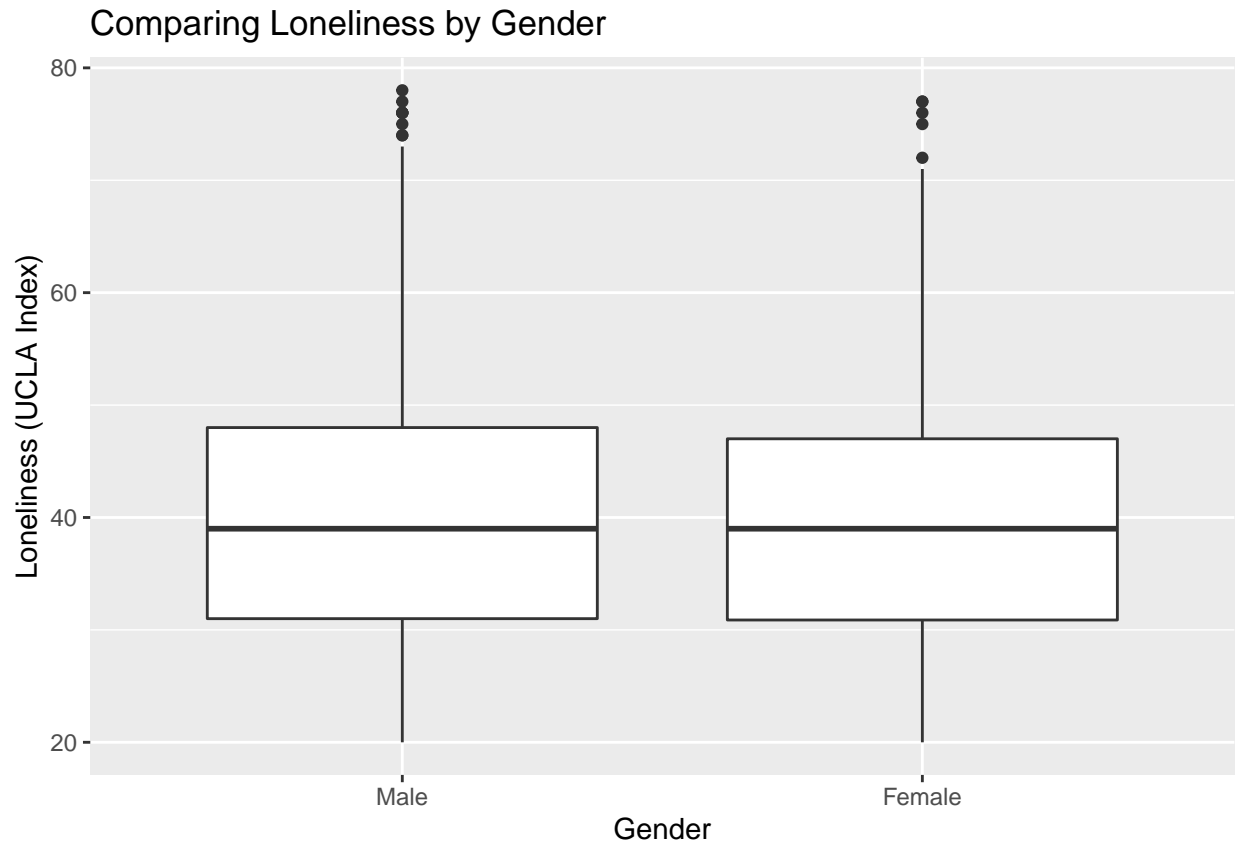
```
#boxplot showing higher level of loneliness when you have a disability
df %>% filter(Q8_disability != "NA") %>%
  ggplot(aes(as.factor(Q8_disability), UCLA_index)) + geom_boxplot() +
  xlab("Do you have a disability?") + ylab("Loneliness (UCLA Index)") +
  labs(title = "Comparing Loneliness by Disability Status") +
  scale_x_discrete(breaks = c("0", "1"),
                  labels = c("No", "Yes"))
```

Comparing Loneliness by Disability Status



#boxplot showing no significant difference in loneliness between genders.

```
df %>% filter(gender != "NA") %>%  
  ggplot(aes(as.factor(gender), UCLA_index)) + geom_boxplot() +  
  xlab("Gender") + ylab("Loneliness (UCLA Index)") +  
  labs(title = "Comparing Loneliness by Gender") +  
  scale_x_discrete(breaks = c("1", "2"),  
                  labels = c("Male", "Female"))
```

My final dataset consisted of 23 predictors, each of which had either a Pearson's correlation coefficient with loneliness of greater than $|0.20|$, significance ($p < .01$) in ANOVA or significance ($p < .05$) in the t-test.

2.2 Modeling Approaches

I started with a linear regression model to see if the 23 variables I selected could out-perform the AARP linear regression model on the basis of R^2 . (It did.)

I then created an XGBoost model with the same variables. I chose XGBoost because I wanted to see if a tree-based model would be more accurate than a linear model, given that most of the linear relationships as measured by Pearson's correlation coefficient were weak, and some of my variables were categorical. I chose XGBoost rather than Random Forest because it is more efficient and accurate. Unlike Random Forest, which creates decision trees in parallel and averages them (for regression), XGBoost uses a *gradient boosting* technique in which decision trees are created sequentially. It increases the weights of variables that were predicted wrong by one tree prior to feeding it into the next tree. The result is often a more accurate model.

Although I reduced the number of predictors from 67 to 23, I still had a large number of variables. I therefore thought a dimension reduction approach would be helpful, and tried principal components regression (PCR.) PCR regresses the principal components of the predictors against the outcome variable rather than using predictors themselves. It can lower the number of parameters in the model. However, because the outcomes are described in terms of principal components rather than variables, it can be more difficult to interpret.

Finally, I created an ensemble as a way to combine multiple weaker models into a stronger model. I computed the ensemble prediction by taking the average of the predictions for the linear regression, XGBoost and PCR models.

I evaluated the relative performance of the linear regression, XGBoost, PCR and ensemble models on the basis of RMSE.

I identified the the variables that were most important to predicting loneliness in the linear regression and XGBoost models. In the regression model, I used the absolute value of the t-score as my indicator of variable importance. The rationale for using t-scores (rather than normalized coefficients) is that it gives us the variables that most certainly have non-zero effects and takes into account the uncertainty in the regression coefficients (“Variable Importance - Linear Regression | Random Effect,” n.d.). In the XGBoost model, I used the Gain metric generated by the `rgb.importance` function. Gain represents the fractional improvement in accuracy that a predictor brings to the branch that it is on. A higher percentage Gain indicates a more important predictor (Abu-Rmieleh 2021).

I examined the top five and top ten most important variables in each model to see where there were commonalities.

3. Models and Results

Before building the models, I split the data 80/20 into training and test sets. All models were built on the training data and RMSE computed in the test set.

```
#Create the relevant data frame with 23 predictor variables and the loneliness variable
variables <- df %>% select(UCLA_index, Q77_hrs_alone, Q27_43_friend_phone,
                          income, Q30_supportive_num, DiversityDiscussImportant,
                          Q89_7_internet_sentiment, Q89_5_internet_sentiment,
                          Q28_discuss_important_matters_num, Q2_health_overall,
                          Q27_41_friend_inperson, NeighborIndex, DiversitySupportive,
                          CSI,
                          ethnic, Q22_marital_satn, Q38_more_less_friends,
                          Q50_groups, Q90_tradeoffs_family,
                          Q91_tradeoffs_intimate_convo, age_group,
                          Q48_volunteer, Q8_disability,
                          Q39_caregiver)

#convert binary and nominal variables to factor class
variables$ethnic <- as.factor(variables$ethnic)
variables$Q22_marital_satn <- as.factor(variables$Q22_marital_satn)
variables$Q38_more_less_friends <- as.factor(variables$Q38_more_less_friends)
variables$Q50_groups <- as.factor(variables$Q50_groups)
variables$Q90_tradeoffs_family <- as.factor(variables$Q90_tradeoffs_family)
variables$Q91_tradeoffs_intimate_convo <- as.factor(variables$Q91_tradeoffs_intimate_convo)
variables$age_group <- as.factor(variables$age_group)
variables$Q48_volunteer <- as.factor(variables$Q48_volunteer)
variables$Q8_disability <- as.factor(variables$Q8_disability)
variables$Q39_caregiver <- as.factor(variables$Q39_caregiver)

#Make sure the outcome variable is in the last column
variables <- variables %>% select(-UCLA_index, UCLA_index)

#convert to a data frame
variables <- as.data.frame(variables)

#Split the data 80/20 into training and test sets.
library(caret)
```

```
set.seed(2, sample.kind="Rounding")
train_index <- createDataPartition(variables$UCLA_index, times = 1, p=.8, list = FALSE)
train <- variables[train_index, ]
test <- variables[-train_index, ]
```

3.1 Linear Regression Model

Below is the code for building the linear regression model, printing the summary with significance levels and R-squared, and computing RMSE in the test set.

```
#Create the linear regression model using the training set
LM <- lm(UCLA_index ~ ., data = train)
```

```
#print out the summary of the model
LMsummary <- summary(LM)
LMsummary
```

```
##
## Call:
## lm(formula = UCLA_index ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6392  -5.6650  -0.0972   5.5826  27.6253
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    64.51084    1.56427   41.240 < 2e-16 ***
## Q77_hrs_alone     1.58246    0.14865   10.645 < 2e-16 ***
## Q27_43_friend_phone -0.37826    0.18983   -1.993  0.046426 *
## income          -0.13048    0.04673   -2.792  0.005279 **
## Q30_supportive_num -0.03546    0.01851   -1.915  0.055564 .
## DiversityDiscussImportant -0.21159    0.22024   -0.961  0.336791
## Q89_7_internet_sentiment -0.65160    0.17933   -3.633  0.000286 ***
## Q89_5_internet_sentiment -1.22255    0.17804   -6.867  8.51e-12 ***
## Q28_discuss_important_matters_num -0.19065    0.08953   -2.129  0.033334 *
## Q2_health_overall  -1.14349    0.22736   -5.029  5.32e-07 ***
## Q27_41_friend_inperson -1.09548    0.21660   -5.058  4.59e-07 ***
## NeighborIndex     -0.52039    0.07956   -6.541  7.59e-11 ***
## DiversitySupportive -2.09363    0.46888   -4.465  8.40e-06 ***
## CSI               1.19288    0.43206    2.761  0.005811 **
## ethnic2           -1.32315    0.63419   -2.086  0.037060 *
## ethnic3            -1.45903    1.11687   -1.306  0.191567
## ethnic4            -1.68767    0.60967   -2.768  0.005684 **
## ethnic5            -0.09572    1.13228   -0.085  0.932640
## Q22_marital_satn1    0.13354    0.89637    0.149  0.881583
## Q22_marital_satn2    4.44113    1.12080    3.962  7.65e-05 ***
## Q22_marital_satn3    3.64916    0.95843    3.807  0.000144 ***
## Q22_marital_satn4    0.63838    0.75040    0.851  0.395021
## Q22_marital_satn5   -2.64805    0.63557   -4.166  3.21e-05 ***
## Q38_more_less_friends2 -3.01616    0.44657   -6.754  1.83e-11 ***
## Q38_more_less_friends3 -3.79740    0.61532   -6.171  8.03e-10 ***
```

```
## Q50_groups2          -1.34922      0.66258  -2.036  0.041839 *
## Q50_groups3          -2.05813      0.82671  -2.490  0.012863 *
## Q50_groups4          -2.46220      1.03116  -2.388  0.017033 *
## Q90_tradeoffs_family2 -1.06453      0.58237  -1.828  0.067695 .
## Q90_tradeoffs_family3 -2.58870      0.81387  -3.181  0.001489 **
## Q91_tradeoffs_intimate_convo2 -2.56839      0.52375  -4.904  1.01e-06 ***
## Q91_tradeoffs_intimate_convo3 -1.70151      0.83620  -2.035  0.041988 *
## age_group6          -0.85156      0.47152  -1.806  0.071057 .
## age_group7          -2.15896      0.54856  -3.936  8.55e-05 ***
## age_group8          -4.40931      0.70868  -6.222  5.86e-10 ***
## Q48_volunteer1      -1.19895      0.61312  -1.955  0.050651 .
## Q8_disability1       1.42819      0.50773   2.813  0.004953 **
## Q39_caregiver1       1.92942      0.50750   3.802  0.000148 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.325 on 2211 degrees of freedom
## (315 observations deleted due to missingness)
## Multiple R-squared:  0.4249, Adjusted R-squared:  0.4153
## F-statistic: 44.15 on 37 and 2211 DF, p-value: < 2.2e-16
```

```
#Create a function to compute RMSE
RMSE <- function(actual_ratings, predicted_ratings){
  sqrt(mean((actual_ratings - predicted_ratings)^2, na.rm = TRUE))
}

#generate predictions on the test set
predict_LM <- predict(LM, newdata = test)

#compute RMSE for the linear regression model
RMSE_LM <- RMSE(test$UCLA_index, predict_LM)
RMSE_LM
```

```
## [1] 7.94992
```

The linear regression model yields an adjusted R-squared of 41.53 and an RMSE in the test set of 7.950. The table below compares these results with the AARP linear regression. The AARP model, which contained over 50 variables yielded an $R^2 = 0.213$. My linear regression model explained almost twice as much of the variance in the outcome variable.

```
#Store the linear regression results in a table and compare
#to the AARP performance
model_results <- tibble(Model = c("AARP", "Linear Regression"),
  Adjusted_R2 = c(.213, LMsummary$adj.r.squared),
  RMSE = c(NA, RMSE_LM))
model_results %>% knitr::kable()
```

Model	Adjusted_R2	RMSE
AARP	0.2130000	NA
Linear Regression	0.4152828	7.94992

3.2 XGBoost Model

Below is the code for fitting and tuning the XGBoost model.

```
#load the xgboost package
library(xgboost)

#Create separate objects for the predictor and outcome variables in the training set
train_x <- data.matrix(train[ ,ncol(train)])
train_y <- train[ , ncol(train)]

#Create separate objects for the predictor and response variables in the test set
test_x <- data.matrix(test[ ,ncol(test)])
test_y <- test[ , ncol(test)]

#Define the final training and testing sets.
xgb_train = xgb.DMatrix(data = train_x, label = train_y)
xgb_test = xgb.DMatrix(data = test_x, label = test_y)

#Fit and tune the model. First define the watchlist
watchlist = list(train=xgb_train, test=xgb_test)

#Fit the XGBoost model and display the training and testing data at each round
set.seed(2, sample.kind="Rounding")
model_XG = xgb.train(data = xgb_train, max.depth = 3, watchlist=watchlist, nrounds = 70)

## [1] train-rmse:29.486989 test-rmse:29.578601
## [2] train-rmse:21.730417 test-rmse:21.851331
## [3] train-rmse:16.577929 test-rmse:16.742274
## [4] train-rmse:13.250493 test-rmse:13.434962
## [5] train-rmse:11.197256 test-rmse:11.446410
## [6] train-rmse:9.962569 test-rmse:10.298171
## [7] train-rmse:9.226861 test-rmse:9.621707
## [8] train-rmse:8.794883 test-rmse:9.223249
## [9] train-rmse:8.531663 test-rmse:8.996041
## [10] train-rmse:8.348206 test-rmse:8.837393
## [11] train-rmse:8.241927 test-rmse:8.760532
## [12] train-rmse:8.152464 test-rmse:8.660068
## [13] train-rmse:8.087599 test-rmse:8.630103
## [14] train-rmse:8.036833 test-rmse:8.603567
## [15] train-rmse:7.996164 test-rmse:8.553246
## [16] train-rmse:7.948341 test-rmse:8.523716
## [17] train-rmse:7.907028 test-rmse:8.497382
## [18] train-rmse:7.859900 test-rmse:8.484594
## [19] train-rmse:7.815695 test-rmse:8.457700
## [20] train-rmse:7.784342 test-rmse:8.448484
## [21] train-rmse:7.755187 test-rmse:8.414996
## [22] train-rmse:7.729497 test-rmse:8.409334
## [23] train-rmse:7.698795 test-rmse:8.396251
## [24] train-rmse:7.675488 test-rmse:8.384122
## [25] train-rmse:7.665717 test-rmse:8.384767
## [26] train-rmse:7.637970 test-rmse:8.385073
## [27] train-rmse:7.611871 test-rmse:8.382819
## [28] train-rmse:7.593398 test-rmse:8.366231
```

```
## [29] train-rmse:7.567768 test-rmse:8.363145
## [30] train-rmse:7.547800 test-rmse:8.374626
## [31] train-rmse:7.536046 test-rmse:8.373139
## [32] train-rmse:7.525187 test-rmse:8.370241
## [33] train-rmse:7.513220 test-rmse:8.354466
## [34] train-rmse:7.479812 test-rmse:8.341227
## [35] train-rmse:7.462732 test-rmse:8.346497
## [36] train-rmse:7.441799 test-rmse:8.346922
## [37] train-rmse:7.433281 test-rmse:8.338707
## [38] train-rmse:7.412193 test-rmse:8.343437
## [39] train-rmse:7.395950 test-rmse:8.352533
## [40] train-rmse:7.375471 test-rmse:8.363573
## [41] train-rmse:7.357914 test-rmse:8.359103
## [42] train-rmse:7.345241 test-rmse:8.352324
## [43] train-rmse:7.329830 test-rmse:8.361366
## [44] train-rmse:7.312186 test-rmse:8.348143
## [45] train-rmse:7.293408 test-rmse:8.316597
## [46] train-rmse:7.278738 test-rmse:8.314274
## [47] train-rmse:7.262073 test-rmse:8.308970
## [48] train-rmse:7.244770 test-rmse:8.297479
## [49] train-rmse:7.233773 test-rmse:8.302850
## [50] train-rmse:7.217058 test-rmse:8.295362
## [51] train-rmse:7.200347 test-rmse:8.290831
## [52] train-rmse:7.186363 test-rmse:8.305541
## [53] train-rmse:7.171670 test-rmse:8.302815
## [54] train-rmse:7.161494 test-rmse:8.303241
## [55] train-rmse:7.148002 test-rmse:8.313640
## [56] train-rmse:7.137662 test-rmse:8.311287
## [57] train-rmse:7.123148 test-rmse:8.318779
## [58] train-rmse:7.105134 test-rmse:8.326676
## [59] train-rmse:7.088669 test-rmse:8.335913
## [60] train-rmse:7.081620 test-rmse:8.339999
## [61] train-rmse:7.063128 test-rmse:8.342937
## [62] train-rmse:7.056711 test-rmse:8.342384
## [63] train-rmse:7.044755 test-rmse:8.350295
## [64] train-rmse:7.034615 test-rmse:8.353435
## [65] train-rmse:7.029626 test-rmse:8.356246
## [66] train-rmse:7.011212 test-rmse:8.350670
## [67] train-rmse:7.002024 test-rmse:8.361233
## [68] train-rmse:6.991727 test-rmse:8.365565
## [69] train-rmse:6.977224 test-rmse:8.355774
## [70] train-rmse:6.966176 test-rmse:8.355578
```

```
#Find the lowest RMSE and insert in the code for nrounds.
set.seed(2, sample.kind="Rounding")
final_model = xgboost(data = xgb_train, max.depth = 3, nrounds = 51, verbose = 0)

#Make predictions on the test set and compute RMSE
predict_XG <- predict(final_model, newdata = xgb_test)
RMSE_XG <- caret::RMSE(test_y, predict_XG)
```

The table below compares the XGBoost model to the linear regression on the basis of RMSE. The XGBoost model, with an RMSE of 8.290 did not perform as well as the linear regression, with an RMSE of 7.950.

```
#Add the XGBoost model to the table
model_results <- bind_rows(model_results,
                           tibble(Model = "XGBoost",
                                   Adjusted_R2 = NA,
                                   RMSE = RMSE_XG))
model_results %>% knitr::kable()
```

Model	Adjusted_R2	RMSE
AARP	0.2130000	NA
Linear Regression	0.4152828	7.949920
XGBoost	NA	8.290831

3.3 Principal Components Regression Model (PCR)

Below is the code for fitting and tuning the PCR model.

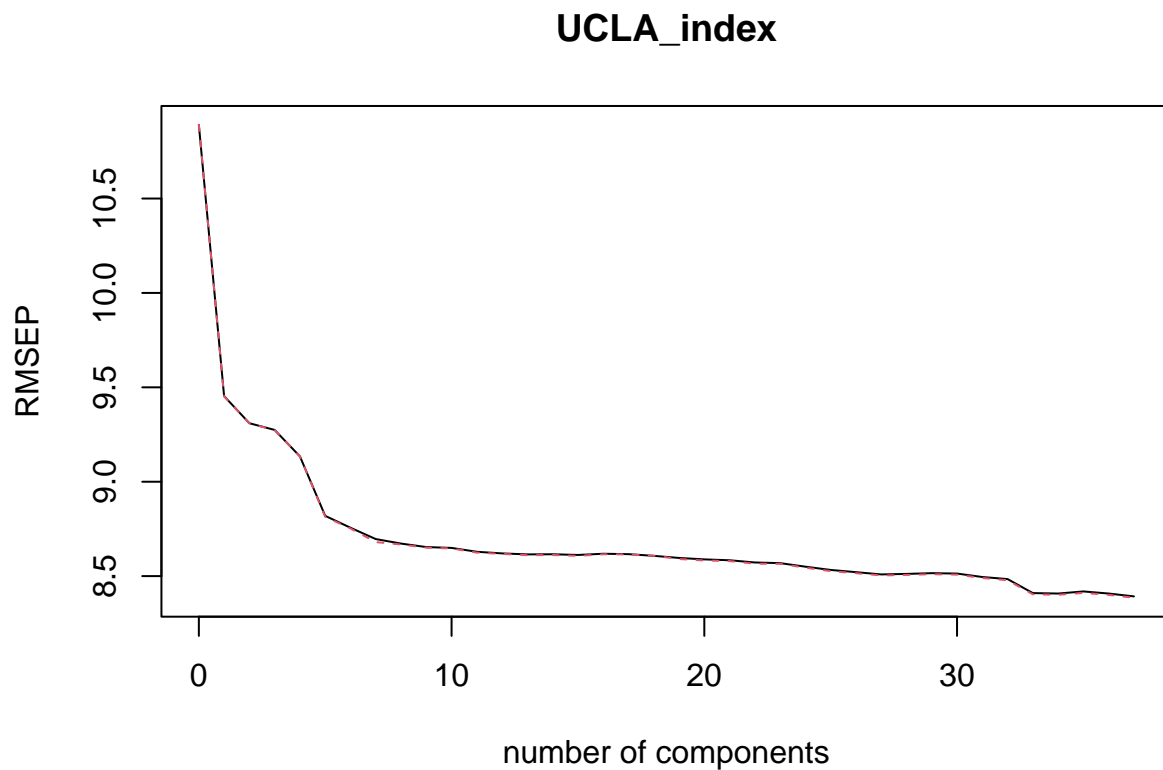
```
#Fit the PCR model
library(pls)
model_PCR <- pcr(UCLA_index ~., data = train, scale=TRUE, validation="CV")

#View a summary of the model fitting and select number of components that yields
#the lowest RMSE in cross-validation
summary(model_PCR)
```

```
## Data:      X dimension: 2249 37
## Y dimension: 2249 1
## Fit method: svdpc
## Number of components considered: 37
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           10.89   9.453   9.310   9.274   9.134   8.819   8.756
## adjCV        10.89   9.452   9.309   9.273   9.134   8.814   8.752
##      7 comps  8 comps  9 comps 10 comps 11 comps 12 comps 13 comps
## CV           8.696   8.672   8.654   8.649   8.629   8.620   8.615
## adjCV        8.680   8.667   8.651   8.647   8.624   8.618   8.611
##      14 comps 15 comps 16 comps 17 comps 18 comps 19 comps 20 comps
## CV           8.616   8.612   8.618   8.616   8.607   8.596   8.589
## adjCV        8.612   8.607   8.616   8.615   8.608   8.591   8.583
##      21 comps 22 comps 23 comps 24 comps 25 comps 26 comps 27 comps
## CV           8.584   8.572   8.569   8.550   8.533   8.521   8.509
## adjCV        8.580   8.566   8.565   8.545   8.526   8.516   8.504
##      28 comps 29 comps 30 comps 31 comps 32 comps 33 comps 34 comps
## CV           8.512   8.516   8.513   8.495   8.484   8.410   8.408
## adjCV        8.506   8.510   8.508   8.490   8.478   8.404   8.401
##      35 comps 36 comps 37 comps
## CV           8.419   8.408   8.393
## adjCV        8.411   8.401   8.385
##
## TRAINING: % variance explained
```

```
##          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X          10.52   17.28   22.41   26.81   30.98   34.87   38.54
## UCLA_index  24.62   26.94   27.59   29.84   34.85   35.68   36.80
##          8 comps  9 comps 10 comps 11 comps 12 comps 13 comps 14 comps
## X          42.12   45.61   48.85   51.98   55.04   58.03   60.88
## UCLA_index  37.06   37.30   37.37   37.74   37.91   38.03   38.04
##          15 comps 16 comps 17 comps 18 comps 19 comps 20 comps
## X          63.67   66.42   69.10   71.74   74.29   76.73
## UCLA_index  38.16   38.16   38.21   38.38   38.83   39.00
##          21 comps 22 comps 23 comps 24 comps 25 comps 26 comps
## X          79.1    81.33   83.47   85.48   87.25   88.93
## UCLA_index  39.0    39.29   39.34   39.66   39.91   40.09
##          27 comps 28 comps 29 comps 30 comps 31 comps 32 comps
## X          90.56   92.09   93.53   94.91   96.02   97.06
## UCLA_index  40.32   40.33   40.34   40.37   40.70   40.94
##          33 comps 34 comps 35 comps 36 comps 37 comps
## X          97.89   98.64   99.29   99.94   100.00
## UCLA_index  41.97   42.04   42.14   42.21   42.49
```

```
#view plot to confirm that 33 components gives the lowest RMSE
validationplot(model_PCR)
```



```
#Use the model to make predictions on the test set
predict_PCR <- predict(model_PCR, test, ncomp=33)
```



```
#Compute RMSE for PCR model in the test set
RMSE_PCR <- RMSE(test$UCLA_index, predict_PCR)
RMSE_PCR
```

```
## [1] 7.982405
```

The table below shows the PCR results compared to linear regression and XGBoost. The PCR model, with an RMSE of 7.982, performed better than the XGBoost but not as well as the linear regression.

```
#Add the PCR results to the table
model_results <- bind_rows(model_results,
                           tibble(Model = "PCR",
                                   Adjusted_R2 = NA,
                                   RMSE = RMSE_PCR))
model_results %>% knitr::kable()
```

Model	Adjusted_R2	RMSE
AARP	0.2130000	NA
Linear Regression	0.4152828	7.949920
XGBoost	NA	8.290831
PCR	NA	7.982405

3.4 Ensemble Model

Below is the code for creating an ensemble from the average of the predictions of the linear regression, XGBoost and PCR models.

```
#Create a dataframe with the predictions from the LM, XGBoost and PCR models
ensemble_preds <- data.frame(LM_preds = predict_LM,
                             XG_preds = predict_XG,
                             PCR_preds = as.vector(predict_PCR),
                             Avg = as.vector((predict_LM + predict_XG +
                                                predict_PCR)/3))

#compute RMSE for the ensemble
RMSE_ensemble <- RMSE(test$UCLA_index, ensemble_preds$Avg)
RMSE_ensemble
```

```
## [1] 7.908848
```

As the table below shows, the ensemble, with an RMSE of 7.91, performed better than the other three models.

```
#Add the ensemble to the table
model_results <- bind_rows(model_results,
                           tibble(Model = "Ensemble",
                                   Adjusted_R2 = NA,
                                   RMSE = RMSE_ensemble))
model_results %>% knitr::kable()
```

Model	Adjusted_R2	RMSE
AARP	0.2130000	NA
Linear Regression	0.4152828	7.949920
XGBoost	NA	8.290831
PCR	NA	7.982405
Ensemble	NA	7.908848

3.5 Variable Importance

I used the linear regression and XGBoost models to provide rankings of how important each of the variables was to predicting loneliness.

First I generated the top five predictors from the linear regression based on absolute value of the t-values.

```
#Extract the absolute value of the t-values from the linear regression model
#summary and convert to a data frame
LM_tvalues <- as.data.frame(LMsummary$coefficients[-1,3])
LM_tvalues <- data.frame(Variable = rownames(LM_tvalues), LM_tvalues)
rownames(LM_tvalues) <- NULL
names(LM_tvalues)[2] <- "t_values"

#Take the absolute value of the t-values and sort highest to lowest
LM_tvalues$t_values <- abs(LM_tvalues$t_values)
LM_tvalues <- arrange(LM_tvalues, desc(t_values))
#Extract the top five predictors from the linear regression model
top_five_LM <- head(LM_tvalues, 5)
top_five_LM
```

```
##           Variable  t_values
## 1      Q77_hrs_alone 10.645220
## 2 Q89_5_internet_sentiment 6.866658
## 3  Q38_more_less_friends2 6.754008
## 4      NeighborIndex 6.540587
## 5      age_group8 6.221880
```

Then I generated the top five predictors from the XGBoost model based the Gain metric in the feature importance matrix.

```
# Generate the feature importance matrix for the XGBoost model
XGimportance_matrix = xgb.importance(colnames(xgb_train), model = final_model)
XGimportance_matrix <- XGimportance_matrix[, 1:2]

#The matrix is already sorted in descending order
#Extract the top five predictors from the XGBoost model
top_five_XG <- head(XGimportance_matrix, 5)
top_five_XG
```

```
##           Feature      Gain
## 1:  Q30_supportive_num 0.16655884
## 2:      NeighborIndex 0.10297016
## 3:      Q77_hrs_alone 0.10080670
## 4: Q89_5_internet_sentiment 0.09071484
## 5:  Q38_more_less_friends 0.07136437
```

Comparing the two lists, we see that they have four factors in common:

- Hours spent physically alone (Q77_hrs_alone)
- Relationships with neighbors (NeighborIndex)
- Agreement with the statement that “the more I use the internet as a replacement for other forms of communication, the lonelier I feel” (Q89_5_internet_sentiment)
- Change in numbers of friends over the last five years. (Q38_more_less_friends)

Thus, these four factors are consistently found to be top predictors of loneliness.

4. Conclusion

4.1 Discussion of Results and Potential Impact

For this project, I created four machine learning models to predict loneliness from a set of social, technology, and demographic factors. The goals were to (a) model the relationships with better performance in terms of R-squared than the AARP model, (b) maximize accuracy in the test set on the basis of RMSE, and (c) identify the features most important to predicting loneliness.

My linear regression model substantially out-performed the AARP model. With an R-squared of 41.5%, my model explained almost double the variability in the loneliness measure as the AARP model, which had an R-squared of 21.3%. Although both of these values may appear low, it is not unusual in the social sciences to see R-squared values less than 50%. It was interesting to see that my model with 23 predictors did better than the AARP model with more than 50 predictors. Typically, adding more variables to a linear regression model increases R-squared. Perhaps greater selectivity reduced redundancy (multicollinearity) or eliminated irrelevant variables, thereby improving the model.

Of the four machine learning models I evaluated on the basis of RMSE, the ensemble performed best. Ensemble methods typically perform better than their constituent models, so this was not surprising. The next best performing model was the linear regression. I had expected XGBoost to perform better than linear regression because it would do a better job of fitting categorical variables, but in the current project it did not offer any improvement. Nor did the PCR. It was notable in the PCR model that the optimal number of components was 33, indicating that it did not offer any dimension reduction.

I examined feature importance using the linear regression and XGBoost models. The two models identified a fairly consistent set of top predictors. Four of the top five were the same for both models:

- Hours spent physically alone
- Relationships with neighbors
- Agreement that using the internet as a replacement for other forms of communication leads to loneliness
- Change in number of friends over the past five years.

Three of these predictors suggest that offline interaction is helpful for avoiding loneliness. Hours spent physically alone were positively associated with loneliness; to avoid hours spent physically alone, we need to engage with others face-to-face rather than online. The NeighborIndex variable includes questions about how often you speak to neighbors and how strong a relationship you have with them. More interaction with neighbors helps to minimize loneliness, and it is conceivable that because neighbors are physically proximate, more of these interactions occur face-to-face. While the question asking respondents whether they agree that *using the internet as a replacement for other forms of communication leads to loneliness* didn’t measure their

actual behavior, it nonetheless suggests that people who recognize the value of offline communications tend to be less lonely.

The current findings and their implications for offline interactions are consistent with prior research exploring how social technologies affect loneliness. According to a review by Nowland et al (Nowland, Necka, and Cacioppo 2017), digital social interaction can reduce loneliness when it is used as a way to enhance existing offline relationships or forge new relationships, but when social technologies are used to displace offline relationships and avoid the messy demands of face-to-face interaction, feelings of loneliness increase. Thus, maintaining an offline social network is necessary for avoiding loneliness. A study by Twenge et al (Twenge, Spitzberg, and Campbell 2019), however, indicates that, at least in teen populations, offline interaction is being eroded by social technologies, with corresponding increases in loneliness. They found that compared to previous generations, today’s teens are spending less face time with friends in activities such as hanging out at the mall, going to parties, dating, riding in cars for fun, or going to the movies, and these reductions correspond to escalating levels of loneliness. In 2017, 39% of 12th graders felt lonely, up from 26% in 2012, when the use of social technologies began to soar. Granted, the Twenge study was performed on teens/young adults and ours examines an older adult population. Studies such as these nonetheless raise the question of how our diminishing face-to-face time affects feelings of loneliness for people of any age.

4.2 Limitations

- The AARP data were collected from a sample of respondents age 45 and above. We therefore need to be cautious about applying the findings to younger populations.
- The AARP survey was not hypothesis driven and because I didn’t design the survey questions, I was unable to directly test hypotheses (e.g., about the relationship between social technology use and loneliness.)
- The AARP survey included ten questions about “internet sentiment”, asking how strongly respondents agreed with a set of statements such, “The more I use the internet as a replacement for other forms of communication, the lonelier I feel.” The survey did not ask about actual behaviors, e.g., whether respondents have in fact used the internet as a replacement for other forms of communication. We are therefore limited in our ability to draw conclusions about the relationships between *behaviors* and loneliness from these ten questions—only about *attitudes* and loneliness.

4.3 Future Work

Additional hypothesis-driven studies should be conducted to better understand the relationships between online and offline interactions and loneliness. Further, because loneliness is most common among teens and young adults (ages ~16 - 24), future work should examine similar factors in this population. The existing literature does not often explore relationships with neighbors as a factor in reducing loneliness. Further research should be conducted to better understand the role and importance of neighborhood relationships.

Appendix A: AARP Variables

Variable Name	Description	Item Type	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30	Item 31	Item 32	Item 33	Item 34	Item 35	Item 36	Item 37	Item 38	Item 39	Item 40	Item 41	Item 42	Item 43	Item 44	Item 45	Item 46	Item 47	Item 48	Item 49	Item 50	Item 51	Item 52	Item 53	Item 54	Item 55	Item 56	Item 57	Item 58	Item 59	Item 60	Item 61	Item 62	Item 63	Item 64	Item 65	Item 66	Item 67	Item 68	Item 69	Item 70	Item 71	Item 72	Item 73	Item 74	Item 75	Item 76	Item 77	Item 78	Item 79	Item 80	Item 81	Item 82	Item 83	Item 84	Item 85	Item 86	Item 87	Item 88	Item 89	Item 90	Item 91	Item 92	Item 93	Item 94	Item 95	Item 96	Item 97	Item 98	Item 99	Item 100	Item 101	Item 102	Item 103	Item 104	Item 105	Item 106	Item 107	Item 108	Item 109	Item 110	Item 111	Item 112	Item 113	Item 114	Item 115	Item 116	Item 117	Item 118	Item 119	Item 120	Item 121	Item 122	Item 123	Item 124	Item 125	Item 126	Item 127	Item 128	Item 129	Item 130	Item 131	Item 132	Item 133	Item 134	Item 135	Item 136	Item 137	Item 138	Item 139	Item 140	Item 141	Item 142	Item 143	Item 144	Item 145	Item 146	Item 147	Item 148	Item 149	Item 150	Item 151	Item 152	Item 153	Item 154	Item 155	Item 156	Item 157	Item 158	Item 159	Item 160	Item 161	Item 162	Item 163	Item 164	Item 165	Item 166	Item 167	Item 168	Item 169	Item 170	Item 171	Item 172	Item 173	Item 174	Item 175	Item 176	Item 177	Item 178	Item 179	Item 180	Item 181	Item 182	Item 183	Item 184	Item 185	Item 186	Item 187	Item 188	Item 189	Item 190	Item 191	Item 192	Item 193	Item 194	Item 195	Item 196	Item 197	Item 198	Item 199	Item 200	Item 201	Item 202	Item 203	Item 204	Item 205	Item 206	Item 207	Item 208	Item 209	Item 210	Item 211	Item 212	Item 213	Item 214	Item 215	Item 216	Item 217	Item 218	Item 219	Item 220	Item 221	Item 222	Item 223	Item 224	Item 225	Item 226	Item 227	Item 228	Item 229	Item 230	Item 231	Item 232	Item 233	Item 234	Item 235	Item 236	Item 237	Item 238	Item 239	Item 240	Item 241	Item 242	Item 243	Item 244	Item 245	Item 246	Item 247	Item 248	Item 249	Item 250	Item 251	Item 252	Item 253	Item 254	Item 255	Item 256	Item 257	Item 258	Item 259	Item 260	Item 261	Item 262	Item 263	Item 264	Item 265	Item 266	Item 267	Item 268	Item 269	Item 270	Item 271	Item 272	Item 273	Item 274	Item 275	Item 276	Item 277	Item 278	Item 279	Item 280	Item 281	Item 282	Item 283	Item 284	Item 285	Item 286	Item 287	Item 288	Item 289	Item 290	Item 291	Item 292	Item 293	Item 294	Item 295	Item 296	Item 297	Item 298	Item 299	Item 300	Item 301	Item 302	Item 303	Item 304	Item 305	Item 306	Item 307	Item 308	Item 309	Item 310	Item 311	Item 312	Item 313	Item 314	Item 315	Item 316	Item 317	Item 318	Item 319	Item 320	Item 321	Item 322	Item 323	Item 324	Item 325	Item 326	Item 327	Item 328	Item 329	Item 330	Item 331	Item 332	Item 333	Item 334	Item 335	Item 336	Item 337	Item 338	Item 339	Item 340	Item 341	Item 342	Item 343	Item 344	Item 345	Item 346	Item 347	Item 348	Item 349	Item 350	Item 351	Item 352	Item 353	Item 354	Item 355	Item 356	Item 357	Item 358	Item 359	Item 360	Item 361	Item 362	Item 363	Item 364	Item 365	Item 366	Item 367	Item 368	Item 369	Item 370	Item 371	Item 372	Item 373	Item 374	Item 375	Item 376	Item 377	Item 378	Item 379	Item 380	Item 381	Item 382	Item 383	Item 384	Item 385	Item 386	Item 387	Item 388	Item 389	Item 390	Item 391	Item 392	Item 393	Item 394	Item 395	Item 396	Item 397	Item 398	Item 399	Item 400	Item 401	Item 402	Item 403	Item 404	Item 405	Item 406	Item 407	Item 408	Item 409	Item 410	Item 411	Item 412	Item 413	Item 414	Item 415	Item 416	Item 417	Item 418	Item 419	Item 420	Item 421	Item 422	Item 423	Item 424	Item 425	Item 426	Item 427	Item 428	Item 429	Item 430	Item 431	Item 432	Item 433	Item 434	Item 435	Item 436	Item 437	Item 438	Item 439	Item 440	Item 441	Item 442	Item 443	Item 444	Item 445	Item 446	Item 447	Item 448	Item 449	Item 450	Item 451	Item 452	Item 453	Item 454	Item 455	Item 456	Item 457	Item 458	Item 459	Item 460	Item 461	Item 462	Item 463	Item 464	Item 465	Item 466	Item 467	Item 468	Item 469	Item 470	Item 471	Item 472	Item 473	Item 474	Item 475	Item 476	Item 477	Item 478	Item 479	Item 480	Item 481	Item 482	Item 483	Item 484	Item 485	Item 486	Item 487	Item 488	Item 489	Item 490	Item 491	Item 492	Item 493	Item 494	Item 495	Item 496	Item 497	Item 498	Item 499	Item 500	Item 501	Item 502	Item 503	Item 504	Item 505	Item 506	Item 507	Item 508	Item 509	Item 510	Item 511	Item 512	Item 513	Item 514	Item 515	Item 516	Item 517	Item 518	Item 519	Item 520	Item 521	Item 522	Item 523	Item 524	Item 525	Item 526	Item 527	Item 528	Item 529	Item 530	Item 531	Item 532	Item 533	Item 534	Item 535	Item 536	Item 537	Item 538	Item 539	Item 540	Item 541	Item 542	Item 543	Item 544	Item 545	Item 546	Item 547	Item 548	Item 549	Item 550	Item 551	Item 552	Item 553	Item 554	Item 555	Item 556	Item 557	Item 558	Item 559	Item 560	Item 561	Item 562	Item 563	Item 564	Item 565	Item 566	Item 567	Item 568	Item 569	Item 570	Item 571	Item 572	Item 573	Item 574	Item 575	Item 576	Item 577	Item 578	Item 579	Item 580	Item 581	Item 582	Item 583	Item 584	Item 585	Item 586	Item 587	Item 588	Item 589	Item 590	Item 591	Item 592	Item 593	Item 594	Item 595	Item 596	Item 597	Item 598	Item 599	Item 600	Item 601	Item 602	Item 603	Item 604	Item 605	Item 606	Item 607	Item 608	Item 609	Item 610	Item 611	Item 612	Item 613	Item 614	Item 615	Item 616	Item 617	Item 618	Item 619	Item 620	Item 621	Item 622	Item 623	Item 624	Item 625	Item 626	Item 627	Item 628	Item 629	Item 630	Item 631	Item 632	Item 633	Item 634	Item 635	Item 636	Item 637	Item 638	Item 639	Item 640	Item 641	Item 642	Item 643	Item 644	Item 645	Item 646	Item 647	Item 648	Item 649	Item 650	Item 651	Item 652	Item 653	Item 654	Item 655	Item 656	Item 657	Item 658	Item 659	Item 660	Item 661	Item 662	Item 663	Item 664	Item 665	Item 666	Item 667	Item 668	Item 669	Item 670	Item 671	Item 672	Item 673	Item 674	Item 675	Item 676	Item 677	Item 678	Item 679	Item 680	Item 681	Item 682	Item 683	Item 684	Item 685	Item 686	Item 687	Item 688	Item 689	Item 690	Item 691	Item 692	Item 693	Item 694	Item 695	Item 696	Item 697	Item 698	Item 699	Item 700	Item 701	Item 702	Item 703	Item 704	Item 705	Item 706	Item 707	Item 708	Item 709	Item 710	Item 711	Item 712	Item 713	Item 714	Item 715	Item 716	Item 717	Item 718	Item 719	Item 720	Item 721	Item 722	Item 723	Item 724	Item 725	Item 726	Item 727	Item 728	Item 729	Item 730	Item 731	Item 732	Item 733	Item 734	Item 735	Item 736	Item 737	Item 738	Item 739	Item 740	Item 741	Item 742	Item 743	Item 744	Item 745	Item 746	Item 747	Item 748	Item 749	Item 750	Item 751	Item 752	Item 753	Item 754	Item 755	Item 756	Item 757	Item 758	Item 759	Item 760	Item 761	Item 762	Item 763	Item 764	Item 765	Item 766	Item 767	Item 768	Item 769	Item 770	Item 771	Item 772	Item 773	Item 774	Item 775	Item 776	Item 777	Item 778	Item 779	Item 780	Item 781	Item 782	Item 783	Item 784	Item 785	Item 786	Item 787	Item 788	Item 789	Item 790	Item 791	Item 792	Item 793	Item 794	Item 795	Item 796	Item 797	Item 798	Item 799	Item 800	Item 801	Item 802	Item 803	Item 804	Item 805	Item 806	Item 807	Item 808	Item 809	Item 810	Item 811	Item 812	Item 813	Item 814	Item 815	Item 816	Item 817	Item 818	Item 819	Item 820	Item 821	Item 822	Item 823	Item 824	Item 825	Item 826	Item 827	Item 828	Item 829	Item 830	Item 831	Item 832	Item 833	Item 834	Item 835	Item 836	Item 837	Item 838	Item 839	Item 840	Item 841	Item 842	Item 843	Item 844	Item 845	Item 846	Item 847	Item 848	Item 849	Item 850	Item 851	Item 852	Item 853	Item 854	Item 855	Item 856	Item 857	Item 858	Item 859	Item 860	Item 861	Item 862	Item 863	Item 864	Item 865	Item 866	Item 867	Item 868	Item 869	Item 870	Item 871	Item 872	Item 873	Item 874	Item 875	Item 876	Item 877	Item 878	Item 879	Item 880	Item 881	Item 882	Item 883	Item 884	Item 885	Item 886	Item 887	Item 888	Item 889	Item 890	Item 891	Item 892	Item 893	Item 894	Item 895	Item 896	Item 897	Item 898	Item 899	Item 900	Item 901	Item 902	Item 903	Item 904	Item 905	Item 906	Item 907	Item 908	Item 909	Item 910	Item 911	Item 912	Item 913	Item 914	Item 915	Item 916	Item 917	Item 918	Item 919	Item 920	Item 921	Item 922	Item 923	Item 924	Item 925	Item 926	Item 927	Item 928	Item 929	Item 930	Item 931	Item 932	Item 933	Item 934	Item 935	Item 936	Item 937	Item 938	Item 939	Item 940	Item 941	Item 942	Item 943	Item 944	Item 945	Item 946	Item 947	Item 948	Item 949	Item 950	Item 951	Item 952	Item 953	Item 954	Item 955	Item 956	Item 957	Item 958	Item 959	Item 960	Item 961	Item 962	Item 963	Item 964	Item 965	Item 966	Item 967	Item 968	Item 969	Item 970	Item 971	Item 972	Item 973	Item 974	Item 975	Item 976	Item 977	Item 978	Item 979	Item 980	Item 981	Item 982	Item 983	Item 984	Item 985	Item 986	Item 987	Item 988	Item 989	Item 990	Item 991	Item 992	Item 993	Item 994	Item 995	Item 996	Item 997	Item 998	Item 999	Item 1000
age_group	Age in category	Categorical	Continuous number ranging 1 to maximum of 100																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															</																																																																																																																																																																																																																																																																																																																																																																																																								

References

- Abu-Rmileh, Amjad. 2021. “Be Careful When Interpreting Your Features Importance in XGBoost!” <https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7>.
- Berkman, L. F., and S. L. Syme. 1979a. “Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents.” *American Journal of Epidemiology* 109 (2): 186–204. <https://doi.org/10.1093/oxfordjournals.aje.a112674>.
- . 1979b. “Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents.” *American Journal of Epidemiology* 109 (2): 186–204. <https://doi.org/10.1093/oxfordjournals.aje.a112674>.
- Berkman, Lisa F., Maria Melchior, Jean-François Chastang, Isabelle Niedhammer, Annette Leclerc, and Marcel Goldberg. 2004. “Social Integration and Mortality: A Prospective Study of French Employees of Electricity of France—Gas of FranceThe GAZEL Cohort.” *American Journal of Epidemiology* 159 (2): 167–74. <https://doi.org/10.1093/aje/kwh020>.
- Brisette, Ian, Sheldon Cohen, and Teresa E. Seeman. 2000. “Measuring Social Integration and Social Networks.” In, 53–85. New York, NY, US: Oxford University Press. <https://doi.org/10.1093/med:psych/9780195126709.003.0003>.
- Cacioppo, John T., and Stephanie Cacioppo. 2014. “Older Adults Reporting Social Isolation or Loneliness Show Poorer Cognitive Function 4 Years Later.” *Evidence-Based Nursing* 17 (2): 59–60. <https://doi.org/10.1136/eb-2013-101379>.
- Cohen, Sheldon, William J Doyle, David P Skoner, S Rabin, and Jack M Gwaltney. 1997a. “Social Ties and Susceptibility to the Common Cold” 277 (24): 5.
- . 1997b. “Social Ties and Susceptibility to the Common Cold” 277 (24): 5.
- . 1997c. “Social Ties and Susceptibility to the Common Cold.” *JAMA* 277 (24): 5.
- “Community Life Survey 2018-19.” n.d. <https://www.gov.uk/government/statistics/community-life-survey-2018-19>.
- Coombs, Bertha. 2020. “Loneliness Is on the Rise and Younger Workers and Social Media Users Feel It Most, Cigna Survey Finds.” <https://www.cnbc.com/2020/01/23/loneliness-is-rising-younger-workers-and-social-media-users-feel-it-most.html>.
- Coyle, Caitlin, and Elizabeth Dugan. 2012. “Social Isolation, Loneliness and Health Among Older Adults.” *Journal of Aging and Health* 24 (December): 1346–63. <https://doi.org/10.1177/0898264312460275>.
- Everson-Rose, Susan A., and Tené T. Lewis. 2005. “Psychosocial factors and cardiovascular diseases.” *Annual Review of Public Health* 26: 469–500. <https://doi.org/10.1146/annurev.publhealth.26.021304.144542>.
- Hawkey, Louise C., Kristen Wroblewski, Till Kaiser, Maike Luhmann, and L. Philip Schumm. 2019. “Are U.S. older adults getting lonelier? Age, period, and cohort differences.” *Psychology and Aging* 34 (8): 1144–57. <https://doi.org/10.1037/pag0000365>.
- Hawkey, Louise, and John Cacioppo. 2010. “Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms.” *Annals of Behavioral Medicine : A Publication of the Society of Behavioral Medicine* 40 (October): 218–27. <https://doi.org/10.1007/s12160-010-9210-8>.
- Holt-Lunstad, Julianne, Timothy B. Smith, Mark Baker, Tyler Harris, and David Stephenson. 2015. “Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review.” *Perspectives on Psychological Science* 10 (2): 227–37. <https://doi.org/10.1177/1745691614568352>.
- Holt-Lunstad, Julianne, Timothy B. Smith, and J. Bradley Layton. 2010a. “Social Relationships and Mortality Risk: A Meta-Analytic Review.” *PLOS Medicine* 7 (7): e1000316. <https://doi.org/10.1371/journal.pmed.1000316>.
- . 2010b. “Social Relationships and Mortality Risk: A Meta-Analytic Review.” *PLOS Medicine* 7 (7): e1000316. <https://doi.org/10.1371/journal.pmed.1000316>.
- Kim, Eric. 2017. “The Brain and Social Connectedness: Global Council on Brain Health Recommendations on Social Engagement and Brain Health,” January.
- Lieberman, Matthew D. 2013. *Social: Why Our Brains Are Wired to Connect*. First Edition. New York: Crown.
- Nowland, Rebecca, Elizabeth Necka, and John Cacioppo. 2017. “Loneliness and Social Internet Use:

- Pathways to Reconnection in a Digital World?” *Perspectives on Psychological Science* 13 (September): 174569161771305. <https://doi.org/10.1177/1745691617713052>.
- Russell, Daniel W. 1996. “UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure.” *Journal of Personality Assessment* 66 (1): 20–40. https://doi.org/10.1207/s15327752jpa6601_2.
- Twenge, Jean M., Brian H. Spitzberg, and W. Keith Campbell. 2019. “Less in-Person Social Interaction with Peers Among U.S. Adolescents in the 21st Century and Links to Loneliness.” *Journal of Social and Personal Relationships* 36 (6): 1892–1913. <https://doi.org/10.1177/0265407519836170>.
- “Variable Importance - Linear Regression | Random Effect.” n.d. <https://randomeffect.net/post/2020/11/01/variable-importance-linear-regression/>.
- Weissbourd, Richard, Milena Batanova, Virginia Lovison, and Eric Torres. n.d. “How the Pandemic Has Deepened an Epidemic of Loneliness and What We Can Do About It,” 13.