Final Project | Part One

EAST 2: Jonathan Gragg, William Johnson, Douglas Wiley

1 San Francisco International Airport Survey

1.1 Introduction

The San Francisco International Airport conducts a yearly comprehensive survey [the survey] of airport guests to rate their satisfaction of facilities, services and amenities. The goal of the survey is to compare the results to previous years and look for areas of and discover new opportunities for improvement. The survey is comprised of sections:

- Flight Information: choice of airline, the destination, reason for traveling.
- Passenger Experience: airport aesthetics, security and safety.
- SFO Website - access and overall usefulness
- Residence - Bay Area, state, country
- Demographic Information - age, gender, income

The survey version under study was taken from 2015.

1.1.1 Part A

1.1.1.1 Research Questions

This research centers around developing an in-depth view of who are the satisfied and unsatisfied customers. The approach is to explore the data without any explicit hypothesis, through the application of data science essentials: collecting, cleansing, exploring, and visualizing the data. Specifically, the research will focus on:

- **Who is satisfied or unsatisfied?** This research will identify if a customer satisfaction proxy can be created from the survey questions.
- **Who are the customers?** This research will explore the customer's demographic data in the survey.
- **What about their flight habits?** Using the flight information in the survey, show the ways customers are intersecting with the airport.

In total this research seeks to narrate a story - in a most literal sense tell the customer journey. This insight will be useful in creating opportunities for improvement at the facility as well as developing future surveys.

1.1.1.2 Exploratory Data Analysis

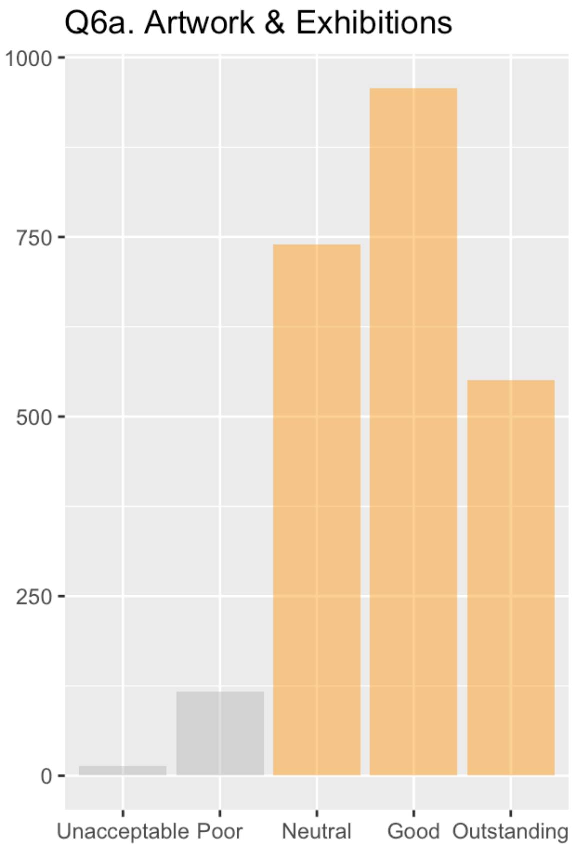
The survey data is cross-sectional, wide format, and attitudinal in nature, implemented using dichotomous (Yes/No), multiple-choice rating questions along with open-ended text. There are 3,234 total observations, with 101 columns (also called features).

Further inspection shows there are a significant number of features with missing values. Overall, this sparsity is not problematic, but most likely by design as these represent question categories without responses as well as empty comment fields.

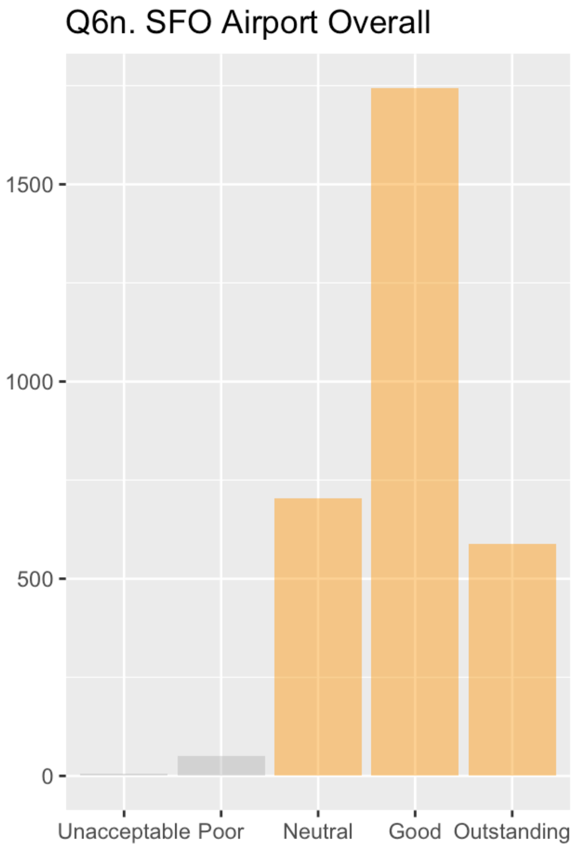
Of primary significance are the results from Question 6. This survey item asks 'How does SFO rate on each of the following attributes?' on 14 categories identified as a, b, c…n. Responses denote a level of acceptability ranging from 1-Unacceptable to 5-Outstanding, with 0 representing a 'blank', and 6 meaning N/A. These ratings will be useful in deriving a sentiment score for each observation. This table summarizes the items for Question 6.

Item	Topic
6a	Artwork and exhibitions
6b	Restaurants
6c	Retail shops and concessions
6d	Signs and directions inside SFO
6e	Escalators/elevators/moving walkways
6f	Information on screens/monitors
6g	Information booths (lower level near baggage claim)
6h	Information booths (upper level – departure area)
6i	Signs and directions on SFO airport roadways
6j	Airport parking facilities
6k	AirTrain
6l	Long term parking lot shuttle
6m	Airport rental car center
6n	SFO Airport as a whole

Shown here are two items from Question 6. These graphics summarize the responses for all of the observations for 6a. Artwork & Exhibitions and 6n. SFO Airport (considering the whole airport).



Q6a. Artwork & Exhibitions



Q6n. SFO Airport Overall

1.1.1.3 Analysis Plan

In answering **who is satisfied or unsatisfied**, the analysis will create a proxy variable 'satisfied' by isolating Questions 6a through 6n. For each observation, use the statistical mode (the value that appears most often) for the question:

- A mode corresponding with either Neutral, Good or Outstanding will result in a positive value.
- A mode corresponding with Poor or Unacceptable will result in a negative value.

Using the 'satisfied' proxy variable from the previous research and survey demographic data, **who are the customers** can be better understood by:

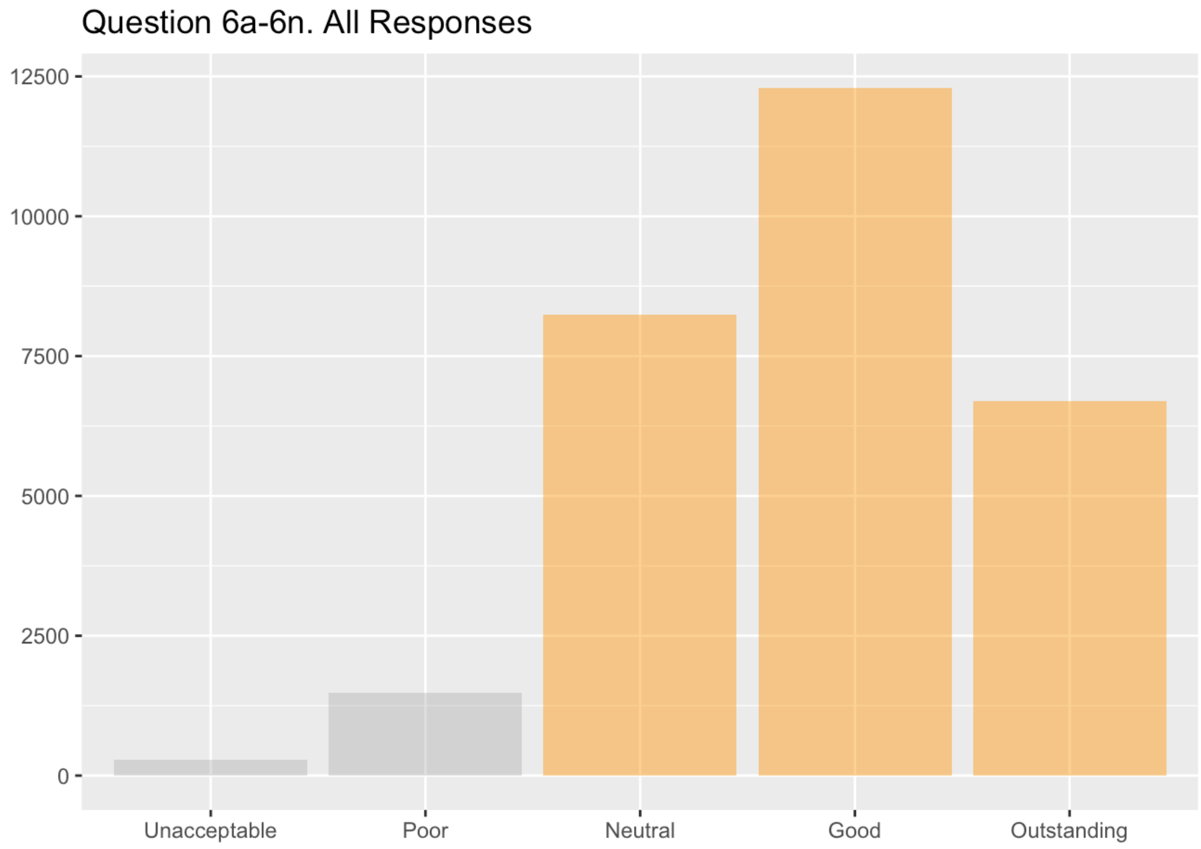
- Plotting the ages (survey question 17) and gender (survey question 18) of the customer and their level of satisfaction.
- Identifying the customer's country of origin (survey question 16) and their level of satisfaction.

Better understanding the **customers flight habits** can be achieved by:

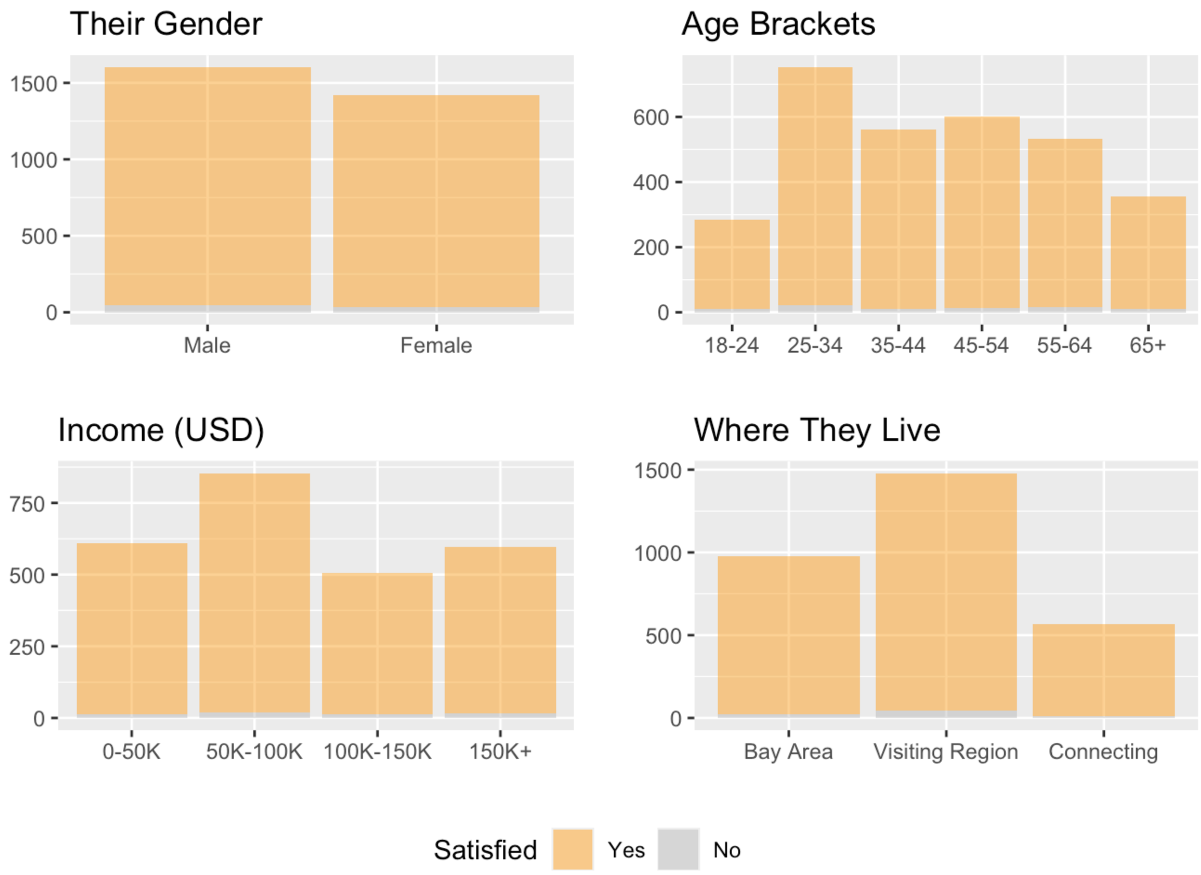
- Using the 'satisfied' proxy variable for the previous research.
- Viewing the customer's yearly number of flights (survey question 5).
- Look at the details of their flight: connection (survey question 1) and destination (survey item destgeo).

1.1.1.4 Results

Survey question 6 was utilized to answer **who is satisfied or unsatisfied**. A look at all responses suggest a normal distribution, thankfully skewed towards more satisfied customers.



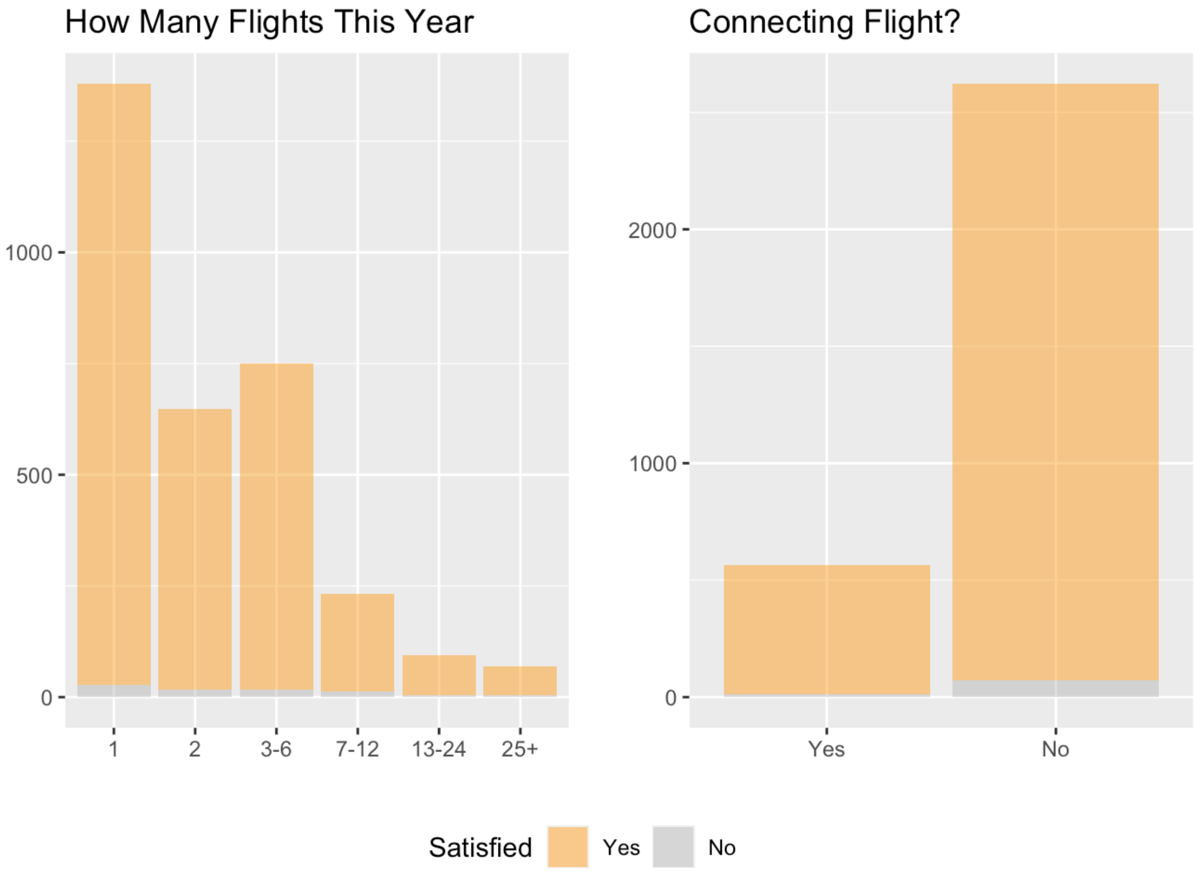
Using the satisfaction proxy created in the previous research, **who are the customers** becomes clear. Survey respondents across all genders, age brackets and levels of income are satisfied with the SFO airport as show here.



Survey respondents come from many different countries. Of the entire survey population, here are the countries that had greater than 20 surveys completed. The good news continues, across all countries:

Country	Surveys	Satisfied
USA	2297	97%
Canada	106	98%
Germany	66	98%
India	61	96%
Japan	57	96%
Australia	56	100%
UK	42	97%
Mexico	41	100%
China	36	100%
New Zealand	23	95%

In viewing **customer flight habits**, for a large majority of survey respondents this was their only flight of the year at the time the survey was recorded. And concurring with the 'Where They Live' chart above, SFO is their final destination - San Francisco is either home or their place to visit.



1.1.1.5 Discussion

The story that is told here is that survey respondents are greatly satisfied with SFO Airport operations. This result is across all genders, ages and income levels. Regardless if the Bay Area is their home or their vacation spot, respondents are giving the airport the highest marks.

For airport leadership, these results can only be viewed as a complete success, but there are problems. The results are heavily positive which ironically hinders the ability to make data-informed decisions. The skewed results may be related to the research itself which was based on a narrowly-defined proxy variable. Though the proxy basis covered wide and important topics, it didn’t factor in other aspects of airport operations. A more sophisticated approach may be to design a composite score that uses the satisfied proxy concept, but include other areas such as security, and sentiment derived from analysis on the textual comments.

Finally, given the imbalances in the results a critical review of the survey itself is required. Does the survey focus too much on range of coverage to the exclusion of specific areas of focus? It may need to be restructured in order to gain more balanced results. Given technological advances in social listening and artificial intelligence, is a survey still the best methodology for understanding the SFO customer? Leveraging these and other innovations will help provide a higher quality of actionable insights for airport leadership.

1.1.1.6 Apendix A: Code

```r
# get source file and convert to dataframe
df_raw <- read.table('SFO_survey_withText.txt', sep='\t', header=T)

# fix column headings
df_raw <- clean_names(df_raw)

# first look
head(df_raw)
```

	respnum	ccgid	term	strata	atype	airline	dest
	<chr>	<chr>	<int>	<int>	<int>	<int>	<int>
1	1	150	1	2	1	25	1
2	2	151	1	2	1	25	1
3	3	152	1	2	1	25	1
4	4	153	1	2	1	25	1
5	5	154	1	2	1	25	1
6	6	155	1	2	1	25	1

6 rows | 1-10 of 102 columns

```
# look at percent missing
sapply(df_raw, function(x) round(((sum(is.na(x)) + sum(is.null(x)) + sum(is.n
an(x)))/length(x)) * 100, 2))
##     respnum       ccgid        term      strata       atype     airline

##        0.00        0.00        0.00        0.00        0.00        0.00

##        dest     gatenum     intdate          q1        q2_1        q2_2

##        0.00        0.00        0.00        0.28        0.00        2.13

##        q2_3        q2_4        q2_5        q2_6        q3_1        q3_2

##       97.09       99.75       99.97       99.97        0.00        0.09

##        q3_3        q3_4        q3_5        q3_6         q3a         q4a

##       99.72       99.97      100.00      100.00        0.71        3.96

##         q4b         q4c          q5       q5avg         saq         q6a

##        8.57        6.93        0.83        0.83        0.00        3.18

##         q6b         q6c         q6d         q6e         q6f         q6g

##        2.50        3.28        2.44        2.29        2.44        3.49

##         q6h         q6i         q6j         q6k         q6l         q6m
```

```
# mcar test not performed as dataset sparsity is possibly by design
#TestMCARNormality(df)
# create a dataframe of q6 responses
df_q6 <- df_raw %>% select(respnum, starts_with('q6'))
```

```
# change q6 to 1...5 with 0 as NA
df_q6[2:15] <- lapply(df_q6[2:15], function(x) ifelse(x==6 | is.null(x) | is.
na(x), 0, x))

# update q6 as an ordered factor
q6_labels <- c("None", "Unacceptable", "Poor", "Neutral", "Good", "Outstandin
g")
df_q6[2:15] <- lapply(df_q6[2:15], function(x) ordered(x, levels = 0:5, label
s = q6_labels))

levels(df_q6$q6a)
## [1] "None"         "Unacceptable" "Poor"         "Neutral"      "Good"

## [6] "Outstanding"
```

```r
q6_items <- data.table(
  Item = c(
    '6a',
    '6b',
    '6c',
    '6d',
    '6e',
    '6f',
    '6g',
    '6h',
    '6i',
    '6j',
    '6k',
    '6l',
    '6m',
    '6n'),
  Topic = c(
    'Artwork and exhibitions',
    'Restaurants',
    'Retail shops and concessions',
    'Signs and directions inside SFO',
    'Escalators/elevators/moving walkways',
    'Information on screens/monitors',
    'Information booths (lower level near baggage claim)',
    'Information booths (upper level – departure area)',
    'Signs and directions on SFO airport roadways',
    'Airport parking facilities',
    'AirTrain',
    'Long term parking lot shuttle',
    'Airport rental car center',
    'SFO Airport as a whole'))

as.htmlwidget(
  formattable(
    q6_items,
    align=c('r', 'l'),
    table.attr = 'class="table table-striped" style="font-size: 11px;"',
    list('Item'=formatter('span', style=~style('font.weight'='bold')))),
  width=500)
```

```
# create a dataframe of q6 responses
df_q6 <- df_raw %>% select(respnum, starts_with('q6'))

# change q6 to 1...5 with 0 as NA
df_q6[2:15] <- lapply(df_q6[2:15], function(x) ifelse(x==6 | is.null(x) | is.
na(x), 0, x))

# update q6 as an ordered factor
q6_labels <- c("None", "Unacceptable", "Poor", "Neutral", "Good", "Outstandin
g")
df_q6[2:15] <- lapply(df_q6[2:15], function(x) ordered(x, levels = 0:5, label
s = q6_labels))

levels(df_q6$q6a)
## [1] "None"         "Unacceptable" "Poor"         "Neutral"      "Good"

## [6] "Outstanding"
```

```r
# eda bar chart #1
plot_q6a <- ggplot(
  data=df_q6 %>% filter(as.integer(q6a) > 1),
  aes(x=q6a)) +
  geom_bar(
    fill=c('gray', 'gray', 'orange', 'orange', 'orange'),
    alpha=0.5,
    na.rm=TRUE) +
    labs(
      title='Q6a. Artwork & Exhibitions',
      x='',
      y='')
# eda bar chart #2
plot_q6n <- ggplot(
   data=df_q6 %>% filter(as.integer(q6n) > 1),
   aes(x=q6n)) +
   geom_bar(
    fill=c('gray', 'gray', 'orange', 'orange', 'orange'),
    alpha=0.5,
    na.rm=TRUE) +
   labs(
     title='Q6n. SFO Airport Overall',
     x='',
     y='')
# use cowplot to assemble into a single graphic
plot_grid_q6a_q6n <- plot_grid(
  plot_q6a + theme(legend.position='none'),
  plot_q6n + theme(legend.position='none'),
  ncol = 2,
  nrow = 1)

# assemble all of the parts
plot_grid(
  plot_grid_q6a_q6n,
  ncol=1,
  rel_heights=c(1))
```

```
# data for chart 3
df_q6_long <- df_q6 %>%
  pivot_longer(cols=all_of(starts_with('q6')), values_to='q6') %>%
  filter(as.integer(q6) > 1)

# eda bar chart #3
ggplot(
  data=df_q6_long,
  aes(x=q6)) +
  geom_bar(
    alpha=0.5,
    fill=c('gray', 'gray', 'orange', 'orange', 'orange'),
    na.rm=TRUE) +
  labs(
    title='Question 6a-6n. All Responses',
    x='',
    y='')
```

```r
# add a satisfied feature to a new df
mode <- function(vals){ which.max(tabulate(vals)) }

# get measures of central tendency for each observation
df_q6_short <- df_q6_long %>%
  filter(as.integer(q6) > 1) %>%
  group_by(respnum) %>%
  summarise(q6_mean = mean(as.integer(q6)),
            q6_mode = mode(as.integer(q6)),
            q6_median = median(as.integer(q6)))

# add a satisfied indicator
df_q6_short <- df_q6_short %>%
  mutate(satisfied = ifelse(q6_mode >= 4, 1, 0))

df_q6_short <- df_q6_short %>%
  mutate(satisfied = ordered(
    satisfied,
    levels=c(1, 0),
    labels=c('Yes', 'No')
  ))

# add the new q6 features back to the df
df <- merge(df_raw, df_q6_short, by='respnum')
```

```r
# create data for gender plot
df_gender <- df %>%
  filter(!is.na(q18)) %>%
  mutate(gender_f=factor(
      q18,
      levels=c(1, 2),
      labels=c('Male', 'Female'))) %>%
  select(respnum, gender=q18, gender_f, satisfied)

# plot of satisfied/unsatisfied by gender
plot_gender <- ggplot(
   data=df_gender,
   aes(
     x=gender_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Their Gender',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))

# create data for age plot
df_age <- df %>%
  filter(!is.na(q17)) %>%
  filter(q17 != 8) %>%
  filter(q17 != 1) %>%
  mutate(age_f=ordered(
      q17,
      levels=c(1, 2, 3, 4, 5, 6, 7),
      labels=c('0-18', '18-24', '25-34', '35-44', '45-54', '55-64', '65+' )))
%>%
  select(respnum, age_f, satisfied)

# plot of satisfied/unsatisfied by age
plot_age <- ggplot(
   data=df_age,
   aes(
     x=age_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Age Brackets',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))
# create data for income plot
df_income <- df %>%
  filter(!is.na(q19)) %>%
  filter(q19 != 5) %>%
  mutate(income_f=ordered(
      q19,
      levels=c(1, 2, 3, 4),
      labels=c('0-50K', '50K-100K', '100K-150K', '150K+'))) %>%
  select(respnum, income_f, satisfied)

# plot of satisfied/unsatisfied by income
plot_income <- ggplot(
   data=df_income,
   aes(
     x=income_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Income (USD)',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))

# create data for residence plot
df_rez <- df %>%
  filter(!is.na(q15)) %>%
  filter(q15 != 4) %>%
  mutate(rez_f=ordered(
      q15,
      levels=c(1, 2, 3),
      labels=c('Bay Area', 'Visiting Region', 'Connecting'))) %>%
  select(respnum, rez_f, satisfied)

# plot of satisfied/unsatisfied by income
plot_rez <- ggplot(
   data=df_rez,
   aes(
     x=rez_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Where They Live',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))
```

```r
# use cowplot to assemble into a single graphic
plot_grid_gender_age <- plot_grid(
  plot_gender + theme(legend.position='none'),
  plot_age    + theme(legend.position='none'),
  plot_income + theme(legend.position='none'),
  plot_rez    + theme(legend.position='none'),
  ncol = 2,
  nrow = 2)

# extract the legend
legend_gender_age <- get_legend(
  plot_gender +
  guides(color = guide_legend(nrow=1)) +
  theme(legend.position='bottom'))
# assemble all of the parts
plot_grid(
  plot_grid_gender_age,
  legend_gender_age,
  ncol=1,
  rel_heights=c(1, 0.1))
```

```
# create data for country plots

levels=c(
  1,2,3,6,
  8,10,12,20,
  21,24,27,29,
  30,32,33,35,
  36,43,44,49,
  52,53,55,56,
  59,60,62)

labels=c(
  'Argentina', 'Australia', 'Austria', 'Belgium',
  'Brazil', 'Canada', 'China', 'France',
  'Germany', 'India', 'Ireland', 'Italy',
  'Japan', 'Korea', 'Mexico', 'Netherlands',
  'New Zealand', 'Peru', 'Philippines', 'Singapore',
  'South Korea', 'Spain', 'Switzerland', 'Taiwan',
  'UAE', 'UK', 'USA')

df_country_percent <-df %>%
  filter(!is.na(country)) %>%
  group_by(country) %>%
  summarize(
    surveys = n(),
    satisfied_percent = as.integer((sum(satisfied=='Yes')/n())*100),
    unsatisfied_percent = as.integer((sum(satisfied=='No')/n())*100)) %>%
  filter(surveys >= 10) %>%
  mutate(country_f=factor(
      country,
      levels=levels,
      labels=labels)) %>%
  select(Country=country_f, Surveys=surveys, Satisfied=satisfied_percent)

as.htmlwidget(
  formattable(
    df_country_percent %>% arrange(desc(Surveys)),
    table.attr = 'class="table table-striped" style="font-size: 11px;"',
    align=c('l', 'r', 'r'),
    list('Country'=formatter('span', style=~style('font.weight'='bold')),
         'Satisfied' = function(x) percent(x/100, digits = 0))),
  width=500)
```

```
# create data for dest plot
df_dest <- df %>%
  filter(!is.na(dest)) %>%
  mutate(dest_f=ordered(
      dest,
      levels=c(1, 2, 3),
      labels=c('Within CA', 'Out of State', 'Out of Country'))) %>%
  select(respnum, dest_f, satisfied)

# plot of satisfied/unsatisfied by destination
plot_dest <- ggplot(
   data=df_dest,
   aes(
     x=dest_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Flight Destination',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))

# create data for connecting plot
df_conn <- df %>%
  filter(!is.na(q1)) %>%
  filter(q1!=0) %>%
  mutate(conn_f=ordered(
      q1,
      levels=c(1, 2),
      labels=c('Yes', 'No'))) %>%
  select(respnum, conn_f, satisfied)

# plot of satisfied/unsatisfied by connection
plot_conn <- ggplot(
   data=df_conn,
   aes(
     x=conn_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='Connecting Flight?',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))
```

```r
# create data for number of flights plot
df_flights <- df %>%
  filter(!is.na(q5)) %>%
  filter(q5 != 0) %>%
  mutate(flights_f=ordered(
      q5,
      levels=c(1, 2, 3, 4, 5, 6),
      labels=c('1', '2', '3-6', '7-12', '13-24', '25+'))) %>%
  select(respnum, flights_f, satisfied)

# plot of satisfied/unsatisfied by connection
plot_flights <- ggplot(
   data=df_flights,
   aes(
     x=flights_f,
     fill=satisfied,
     na.rm=TRUE)) +
   geom_bar(
     alpha = 0.5,
     na.rm=TRUE) +
   labs(
     title='How Many Flights This Year',
     x = '',
     y = '',
     fill='Satisfied') +
   scale_fill_manual(values = c('orange', 'grey'))

# use cowplot to assemble into a single graphic
plot_grid_conn_dest <- plot_grid(
  plot_flights + theme(legend.position='none'),
  plot_conn + theme(legend.position='none'),
#  plot_dest + theme(legend.position='none'),
  ncol = 2,
  nrow = 1)

# extract the legend
legend_gender_conn <- get_legend(
  plot_conn +
  guides(color = guide_legend(nrow=1)) +
  theme(legend.position='bottom'))
# assemble all of the parts
plot_grid(
  plot_grid_conn_dest,
  legend_gender_conn,
  ncol=1,
  rel_heights=c(1, 0.1))
```