# Analyzing Survey Data from the San Francisco Museum of Modern Art

Apoorva Dudani

2022-12-27

**In this report, I analyze survey data collected from 600 museum visitors at the San Francisco Museum of Modern Art (SFMOMA) to answer 7 research questions. Please read the dataset dictionary, which has been attached as a .csv file in the repository, before proceeding with this report.**

**I use a significance level (alpha) = .05 for all statistical tests I conduct.**

First, I load the libraries that I will use in my analyses

```
library("psych")
library("ggplot2")
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

Then, I read in the data file and generate a list of the variables contained in the file.

```
sfmoma.df = read.csv("/Users/Archana/Desktop/usfca/github projects/project 1/sfmoma_dataset.csv", heade
str(sfmoma.df)
```

```
## 'data.frame':    600 obs. of  7 variables:
##  $ cust.id       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ visitalone    : chr  "yesalone" "notalone" "notalone" "notalone" ...
##  $ purchaseticket: chr  "yespurchase" "yespurchase" "yespurchase" "nopurchase" ...
##  $ dollarsspent  : num  37 45.1 42.5 45.7 41.8 ...
##  $ cafedining    : chr  "yesdining" "nodining" "nodining" "nodining" ...
##  $ satisfaction  : int  104 132 127 138 122 151 143 115 113 117 ...
##  $ timespentFB   : int  224 266 253 276 246 302 274 245 232 234 ...
```

For this data set,

visitalone is a binary categorical variable; R reads this in as a character variable. purchaseticket is a binary categorical variable; R reads this in as a character variable. dollarsspent is numeric; R reads in as numeric.

cafedining is a binary categorical variable; R reads this in as a character variable. satisfactcion is numeric; R reads in as integer; this is ok. timespentFB is numeric; R reads in as integer; this is ok.

By default, R changes character data to factors (some R functions require that categorical variables be declared as factors). Hence, I do this coding for the categorical variables:

```
sfmoma.df$visitalone  <- factor(sfmoma.df$visitalone,
                        levels=c("yesalone","notalone"),
                        labels=c("yesalone","notalone"))

sfmoma.df$purchaseticket <- factor(sfmoma.df$purchaseticket,
                            levels=c("yespurchase","nopurchase"),
                            labels=c("yespurchase","nopurchase"))

sfmoma.df$cafedining <- factor(sfmoma.df$cafedining,
                            levels=c("yesdining","nodining"),
                            labels=c("yesdining","nodining"))
```

# Research Question 1

## Are visitors who purchase a ticket to a special exhibit also likely to dine in the cafe?

The variables to analyze are purchaseticket and cafedining, and I want to know if these two variables are related. Both variables are categorical (specifically, both are binary), so I use a chi-square test

*Null hypothesis: purchaseticket and cafedining are independent*

*Alternative hypothesis: purchaseticket and cafedining are dependent*

First, I create a contingency table to display the data.

```
con<-table(sfmoma.df$purchaseticket, sfmoma.df$cafedining)
con
```

```
##
##                yesdining nodining
##   yespurchase        269      205
##   nopurchase          16      110
```

Next, I perform a chi-square test

```
test <- chisq.test(table(sfmoma.df$purchaseticket, sfmoma.df$cafedining))
test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(sfmoma.df$purchaseticket, sfmoma.df$cafedining)
## X-squared = 75.706, df = 1, p-value < 2.2e-16
```

X-squared = 75.706, df = 1, p-value < 2.2e-16

The p-value is smaller than alpha, so I reject the null hypothesis and conclude that purchaseticket and cafedining are dependent.

It's helpful to print the observed and expected counts so that the interpretation is clear.

```
test$observed
```

```
##
##               yesdining nodining
##   yespurchase       269      205
##   nopurchase         16      110
```

```
test$expected
```

```
##
##               yesdining nodining
##   yespurchase    225.15   248.85
##   nopurchase      59.85    66.15
```

Relative to the expected counts, I see that there are more observed counts in the yesdining/yespurchase and nodining/nopurchase cells. Thus, there is a tendency for those who purchase a ticket to a special exhibit to also dine in the cafe.

### Executive summary: Visitors who purchase a ticket to a special exhibit are also likely to dine in the cafe. I recommend that the SFMOMA consider a promotion to encourage visitors to purchase a ticket to special art exhibits to help increase dining in the cafe.

## Research Question 2

### For visitors who do not come alone to the museum, are those who purchase a ticket for a special art exhibit also likely to dine in the cafe?

*Null hypothesis: For visitors who do not come alone to the museum, purchaseticket and cafedining are independent*

*Alternative hypothesis: For visitors who do not come alone to the museum, purchaseticket and cafedining are dependent*

I first need to subset observations in the data frame so that I analyze data for only those visitors who do not come alone to the museum. Here, I've called the new subset of data 'subset_notalone'

```
subset_notalone <- sfmoma.df[sfmoma.df$visitalone == 'notalone',]
test <- chisq.test(table(subset_notalone$purchaseticket, subset_notalone$cafedining))
test
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(subset_notalone$purchaseticket, subset_notalone$cafedining)
## X-squared = 63.344, df = 1, p-value = 1.736e-15
```

X-squared = 63.344, df = 1, p-value = 1.736e-15

The p-value < alpha so I do not reject the null hypothesis, and conclude that for those who do not visit the museum alone, purchaseticket and cafedining are dependent

Again, it's helpful to print the observed and expected counts so that the interpretation is clear.

```
test$observed
```

```
##
##               yesdining nodining
##   yespurchase       198      166
##   nopurchase         13      101
```

```
test$expected
```

```
##
##               yesdining  nodining
##   yespurchase 160.67782 203.32218
##   nopurchase   50.32218  63.67782
```

Relative to the expected counts, I see that there are more observed counts in the yesdining/yespurchase and nodining/nopurchase cells. Thus, there is a tendency for those who purchase a ticket to a special exhibit to also dine in the cafe.

**Executive summary: For visitors who do not visit alone, those who purchase a ticket to a special exhibit are also likely to dine in the cafe. I recommend that SFMOMA consider a promotion to encourage visitors who come in groups of two or more to purchase a ticket to special art exhibits to help increase dining in the cafe.**

# Research Question 3

## Is there a relationship between the amount spent at the museum gift shop and satisfaction with the visit to the museum?

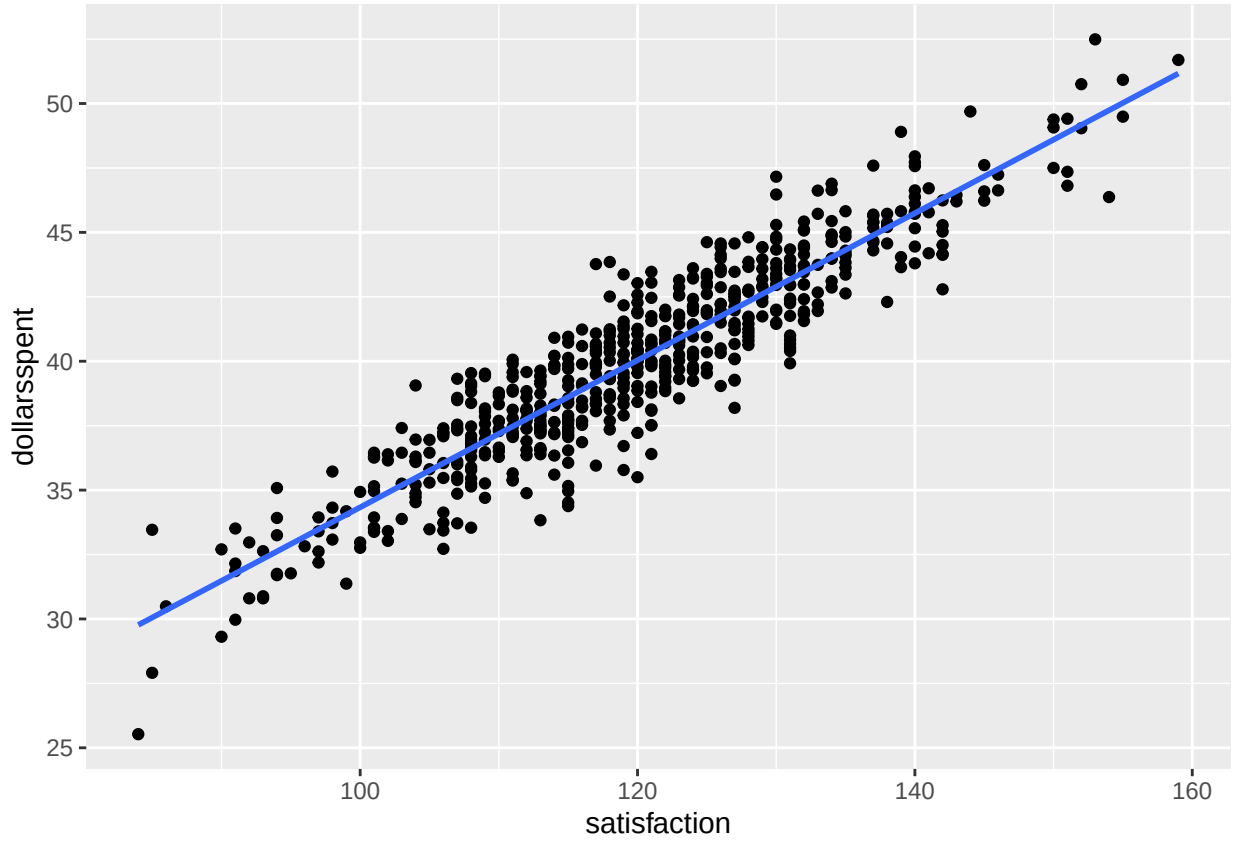
Variables to analyze are dollarsspent and satisfaction, and I want to know if these two variables are related.

Both variables are numeric, so I use correlation, r

First, I create a display of the association between dollarsspent and satisfaction.

```
ggplot(data=sfmoma.df, aes(x=satisfaction, y=dollarsspent)) + geom_point() + geom_smooth(method=lm, se=
```

```
## `geom_smooth()` using formula 'y ~ x'
```




*Null hypothesis: dollarsspent and satisfaction are independent*

*Alternative hypothesis: dollarsspent and satisfaction are dependent*

Now, I calculate the correlation coefficient and carry out a test of significance.

```
corr.test(x=sfmoma.df$satisfaction, y=sfmoma.df$dollarsspent, method="pearson", alpha=.05)
```

```
## Call:corr.test(x = sfmoma.df$satisfaction, y = sfmoma.df$dollarsspent,
##     method = "pearson", alpha = 0.05)
## Correlation matrix
## [1] 0.92
## Sample Size
## [1] 600
## These are the unadjusted probability values.
##   The probability values  adjusted for multiple tests are in the p.adj object.
## [1] 0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

The estimated correlation is 0.92. This is not surprising given the scatterplot produced above suggested a strong relationship between the two variables.

The p-value is 0. From this, I reject the null hypothesis. We conclude that there is a relationship between the two variables.

**Executive summary: There is a relationship between satisfaction with the visit and the dollars spent at the gift shop. Any effort that helps to improve visitor satisfaction might help to increase the amount spent at the gift shop.**

## Research Question 4

### For visitors who do not visit alone, is there a relationship between the amount spent at the museum gift shop and satisfaction with the visit to the museum?

*Null hypothesis: For visitors who do not visit alone, dollarsspent and satisfaction are independent*

*Alternative hypothesis: For visitors who do not visit alone, dollarsspent and satisfaction are dependent*

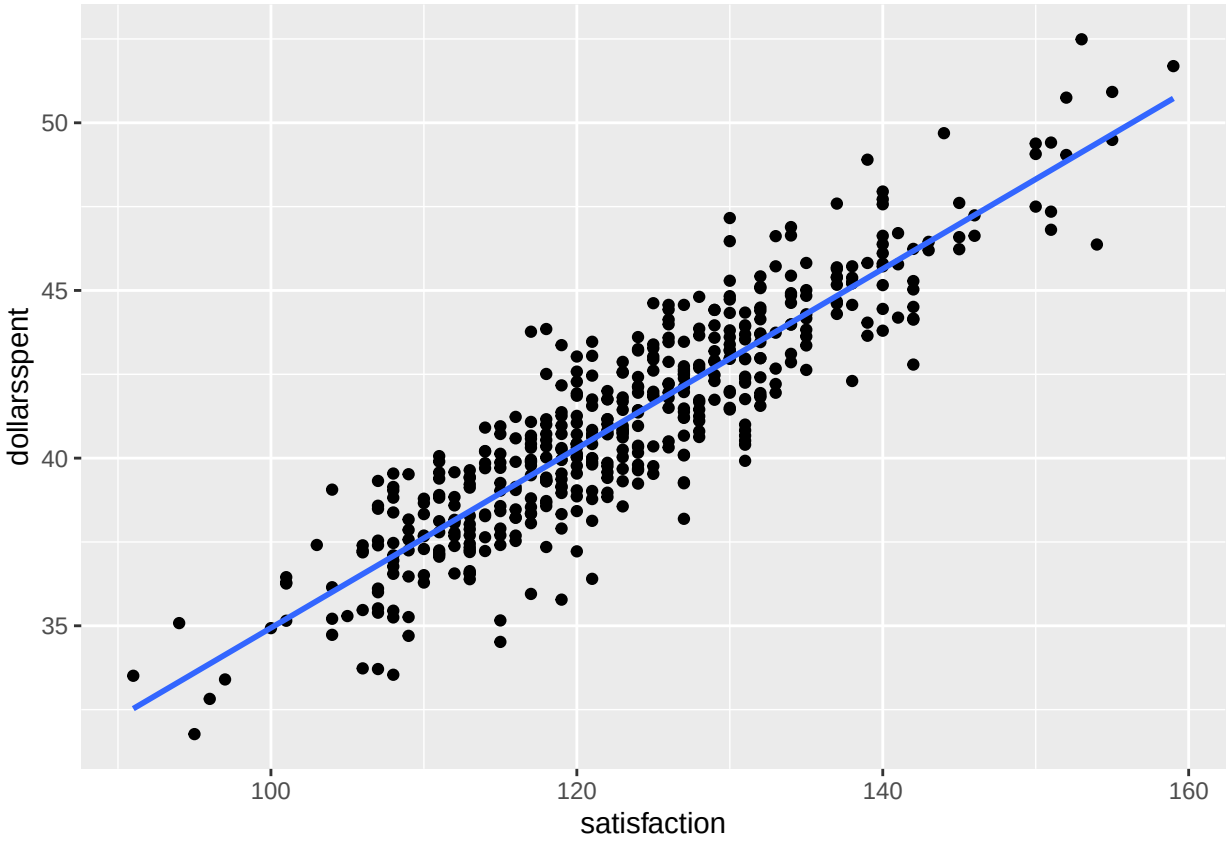
I first subset observations in the data frame so that I analyze data for only those visitors who do not come alone to the museum. Here, I've called the new subset of data 'subset_notalone'

```
subset_notalone <- sfmoma.df[sfmoma.df$visitalone == 'notalone',]
```

Then, I create a display of the association between dollarsspent and satisfaction for visitors who do not visit alone

```
ggplot(data=subset_notalone , aes(x=satisfaction, y=dollarsspent)) + geom_point() + geom_smooth(method=l
```

```
## `geom_smooth()` using formula 'y ~ x'
```




Now, I calculate the correlation coefficient and carry out a test of significance.

```
corr.test(x=subset_notalone$satisfaction, y=subset_notalone$dollarsspent, method="pearson", alpha=.05)
```

```
## Call:corr.test(x = subset_notalone$satisfaction, y = subset_notalone$dollarsspent,
##     method = "pearson", alpha = 0.05)
## Correlation matrix
## [1] 0.9
## Sample Size
## [1] 478
## These are the unadjusted probability values.
##   The probability values  adjusted for multiple tests are in the p.adj object.
## [1] 0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

The estimated correlation is .90. This is also not surprising given the scatterplot produced above suggested a strong relationship between the two variables.

The p-value is 0. From this, I reject the null hypothesis and conclude that there is a relationship between the two variables for visitors who do not visit alone.

**Executive summary: For visitors who do not visit alone, satisfaction is related to how much they spend at the gift shop. Any effort to improve satisfaction for visitors who come in groups of two or more may help to increase spending at the gift shop.**

## Research Question 5

### What is the estimated mean satisfaction level according to whether or not a visitor comes alone or with others? Is there an important difference in mean satisfaction between these groups?

The variables I analyze for this research question are satisfaction and visitalone. I want to know if these two variables are related.

visitalone is categorical (specifically, binary), and satisfaction is numeric, so I use a t-test to compare mean satisfaction level between the two groups of customers (those who come alone and those who do not).

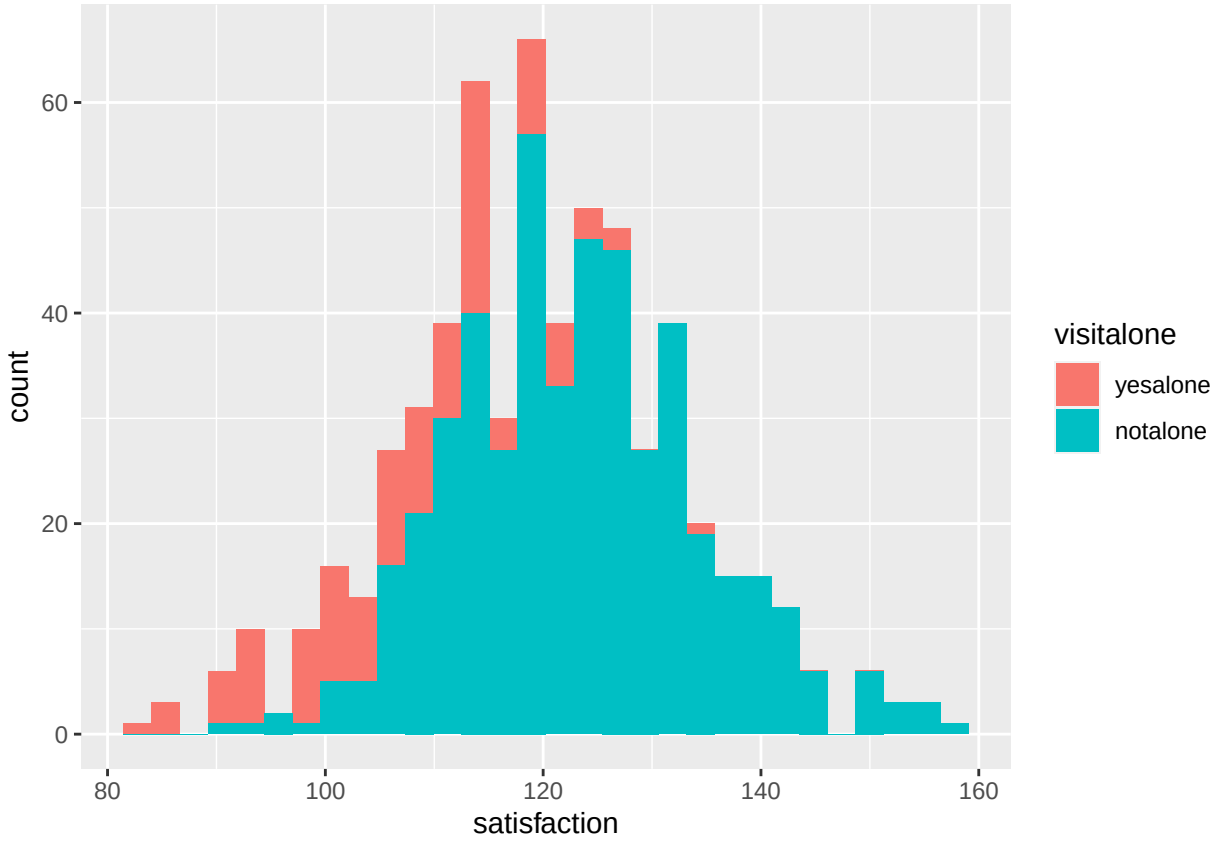
*Null hypothesis: mean satisfaction is equal between the two visitalone groups (visitalone and satisfaction are independent)*

*Alternative hypothesis: mean satisfaction differs between the two visitalone groups (visitalone and satisfaction are dependent)*

Here is a histogram that displays the numeric variable (satisfaction) separately for each of the two levels of the categorical variable (visitalone).

```
ggplot(sfmoma.df, aes(x = satisfaction, fill = visitalone)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




A t-test requires either that each sample size is >30 OR that each population distribution is normal.

```
n_visitalone <- table(sfmoma.df$visitalone)
n_visitalone
```

```
##
## yesalone notalone
##      122      478
```

Each of the two 'visitalone' groups have sample sizes >30, so the t-test is valid.

```
ttest_satisfaction <-t.test(satisfaction ~ visitalone, data=sfmoma.df, var.equal=FALSE)
ttest_satisfaction
```

```
##
##  Welch Two Sample t-test
##
## data:  satisfaction by visitalone
## t = -14.658, df = 201.37, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group yesalone and group notalone is not eq
## 95 percent confidence interval:
##  -17.69624 -13.49966
## sample estimates:
## mean in group yesalone mean in group notalone
##               107.5820               123.1799
```

The p-value is smaller than alpha (0.05), so I reject the null hypothesis. I conclude that satisfaction and visitalone are dependent, and that there is a difference in mean overall satisfaction between those who come alone and those who do not.

**Executive summary: Customer satisfaction is related to whether a visitor comes to the museum alone or with others. Any efforts to improve customer satisfaction should take into account whether or not a customer visits the museum alone. This can include**

# Research Question 6

## What is the estimated mean time spent on Facebook according to whether a visitor comes alone or comes with others? Is there an important mean difference in time spent on Facebook between these groups?

Variables to analyze are timespentFB and visitalone, and I want to know if these two variables are related.

visitalone is categorical (specifically, binary), and timespentFB is numeric, so we'll use a t test to compare satisfaction between the two groups of visitors (those who visit alone and those who visit with others)

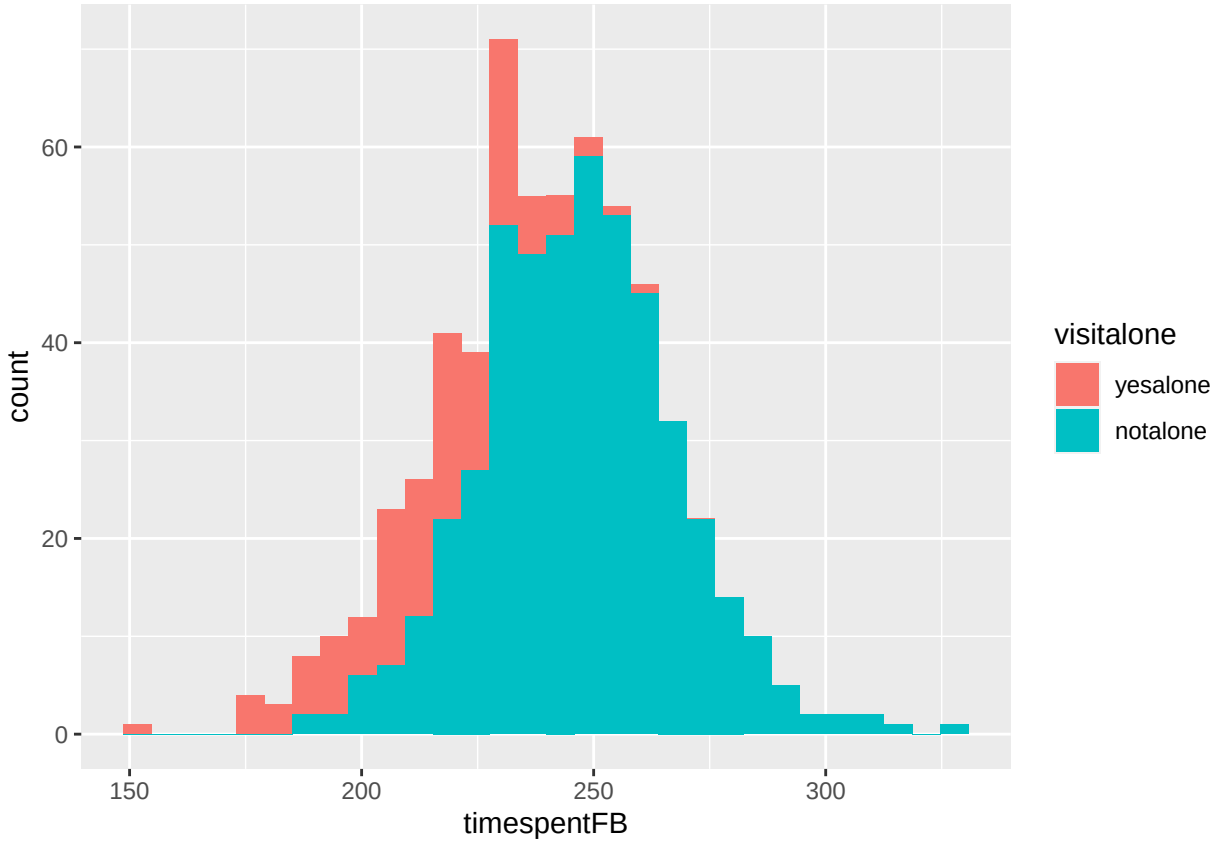
*Null hypothesis: mean timespentFB is equal between the two visitor groups (timespentFB and visitalone are independent)*

*Alternative hypothesis: mean timespentFB differs between the two visitor groups (timespentFB and visitalone are dependent)*

First, I create a histogram that displays the numeric variable (timespentFB) separately for each of the two levels of the categorical variable (visitalone).

```
ggplot(sfmoma.df, aes(x = timespentFB, fill = visitalone)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




A t-test requires either that each sample size is >30 OR that each population distribution is normal.

```
n_visitalone <- table(sfmoma.df$visitalone)
n_visitalone
```

```
##
## yesalone notalone
##      122      478
```

Each of the two 'visitalone' groups have sample sizes >30, so the t test is valid.

I also carry out the Shapiro-Wilk normality test, which evaluates the null hypothesis that the population distribution is normal vs. the alternative hypothesis that the population distribution is not normal.

```
with(sfmoma.df, shapiro.test(timespentFB[visitalone == "yesalone"]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  timespentFB[visitalone == "yesalone"]
## W = 0.98583, p-value = 0.2338
```

```
with(sfmoma.df, shapiro.test(timespentFB[visitalone == "notalone"]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  timespentFB[visitalone == "notalone"]
## W = 0.99426, p-value = 0.0692
```

Above, I carried out two tests of normality: One for the distribution of timespentFB for those who visit alone and one test for those who do not visit alone. The p-values from the two tests are $> .05$, so I do not reject the null hypothesis and conclude that it's reasonable to assume normally distributed populations.

Now I can carry out a t-test to test the hypotheses

```
ttest_timespentFB  <-t.test(timespentFB ~ visitalone, data=sfmoma.df, var.equal=FALSE)
print(ttest_timespentFB)
```

```
##
##  Welch Two Sample t-test
##
## data:  timespentFB by visitalone
## t = -16.547, df = 208.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group yesalone and group notalone is not eq
## 95 percent confidence interval:
##  -35.65926 -28.06719
## sample estimates:
## mean in group yesalone mean in group notalone
##               214.7623               246.6255
```

From the test, t = -16.547, df = 208.8, p-value $<$ 2.2e-16.

The p-value is $<$ alpha so I reject the null hypothesis and conclude that there is a difference in mean timespentFB between those who visit alone and those who do not.

**Executive summary: Time spent using Facebook is related to whether or not a visitor comes alone or with others. I recommend that if SFMOMA wishes to encourage visits by groups of two or more people that they advertise on Facebook because those visitors tend to spend more time using the platform.**

# Research Question 7

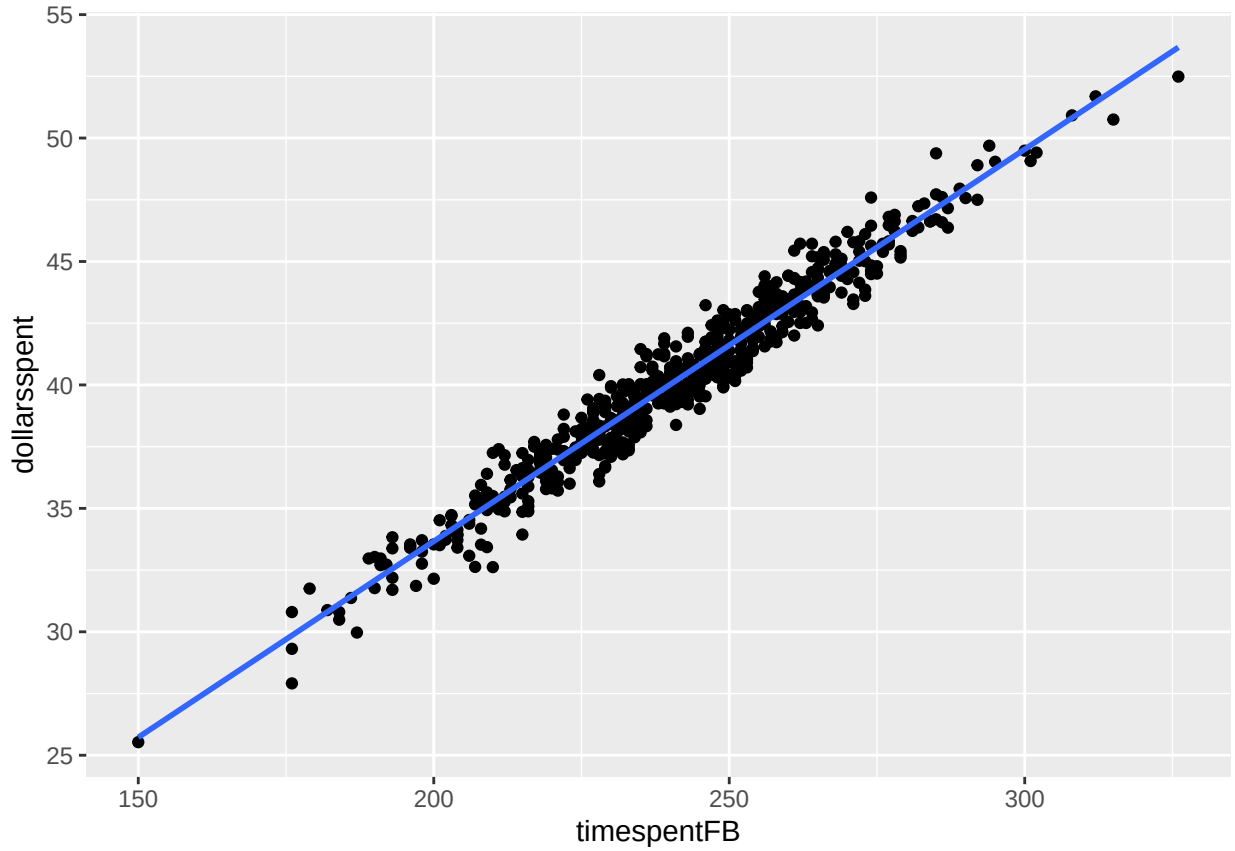
## Is there a relationship between time spent on Facebook and amount spent at the gift shop?

The variables to analyze are timespentFB and dollarsspent, and I want to know if these two variables are related. Both variables are numeric, so I use correlation, r

First, I create a display of the association between dollarsspent and timespentFB

```
ggplot(data=sfmoma.df, aes(x=timespentFB, y=dollarsspent)) + geom_point() + geom_smooth(method=lm, se=F
```

```
## `geom_smooth()` using formula 'y ~ x'
```




*Null hypothesis: dollarsspent and timespentFB are independent*

*Alternative hypothesis: dollarsspent and timespentFB are dependent*

Now, I calculate the correlation coefficient and carry out a test of significance.

```
corr.test(x=sfmoma.df$timespentFB, y=sfmoma.df$dollarsspent, method="pearson", alpha=.05)
```

```
## Call:corr.test(x = sfmoma.df$timespentFB, y = sfmoma.df$dollarsspent,
##     method = "pearson", alpha = 0.05)
## Correlation matrix
## [1] 0.98
## Sample Size
## [1] 600
## These are the unadjusted probability values.
##   The probability values  adjusted for multiple tests are in the p.adj object.
## [1] 0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

The estimated correlation is 0.98. This is not surprising given the scatterplot produced above suggested a very strong relationship between the two variables.

The p-value is 0. From this, I reject the null hypothesis and conclude that there is a relationship between the two variables.

**Executive summary: There is a relationship between time spent on Facebook and the amount spent at the gift shop. To help increase the amount spent at the gift shop, we might recommend that the client advertise on Facebook because those who spend more at the gift shop tend to spend more time using Facebook.**