Analyzing Survey Data from a Prepared Meal Delivery Service Company

Apoorva Dudani

2022-12-30

In this report, I analyze data collected by a local company that prepares meals and ships them to the customer's home. Their marketing group sent out a survey to their customer list and received responses from 600 current customers.

The survey included questions about a customer's age ('age'), annual income (in tens of thousands of dollars) ('annual_income'), typical daily commute time (in minutes)('commute_time'), a scale that measured their openness to new experiences ('open_personality'), diet preference ('diet'), whether the receive meals need assembling ('assemble_meals'), whether they live with others ('live_with_others'), whether they need a gluten-free diet ('gluten_free'), and finally, a scale that measured customer satisfaction with the product ('satisfaction').

I use these to answer two research questions, and use a significance level (alpha) = .05 for all the statistical tests I conduct.

```
library(ggplot2)
library(afex)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
## ************
## Welcome to afex. For support visit: http://afex.singmann.science/
```

```
## - Functions for ANOVAs: aov_car(), aov_ez(), and aov_4()
## - Methods for calculating p-values with mixed(): 'S', 'KR', 'LRT', and 'PB'
## - 'afex_aov' and 'mixed' objects can be passed to emmeans() for follow-up tests
## - NEWS: emmeans() for ANOVA models now uses model = 'multivariate' as default.
## - Get and set global package options with: afex_options()
## - Set orthogonal sum-to-zero contrasts globally: set_sum_contrasts()
## - For example analyses see: browseVignettes("afex")
## ************
```

```
##
## Attaching package: 'afex'

## The following object is masked from 'package:lme4':
##
##     lmer
```

```
library(emmeans)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mealdelivery.df = read.csv("/Users/Archana/Desktop/usfca/github projects/project 2/project 2.csv", heade
str(mealdelivery.df)
```

```
## 'data.frame':    600 obs. of  10 variables:
##  $ cust.id         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ assemble_meals  : chr  "yes" "no" "no" "no" ...
##  $ live_with_others: chr  "no" "yes" "no" "yes" ...
##  $ gluten_free     : chr  "no" "no" "no" "no" ...
##  $ diet            : chr  "AllFoodTypes" "Vegan" "AllFoodTypes" "AllFoodTypes" ...
##  $ satisfaction    : int  39 40 41 38 42 43 39 40 41 39 ...
##  $ age             : int  26 26 31 31 28 35 29 32 34 32 ...
##  $ annual_income   : int  76 67 82 94 67 67 92 63 52 77 ...
##  $ commute_time    : int  71 69 76 82 63 59 68 65 66 73 ...
##  $ open_personality: int  9 11 11 11 9 11 11 9 10 9 ...
```

I see that there are 4 variables listed as character variables. I check the number of levels of each of these character variables.

```
table(mealdelivery.df$assemble_meals)
```

```
##
##  no yes
## 372 228
```

```
table(mealdelivery.df$live_with_others)
```

```
##
##  no yes
## 184 416
```

```
table(mealdelivery.df$gluten_free)
```

```
##
##  no yes
## 475 125
```

```
table(mealdelivery.df$diet)
```

```
##
## AllFoodTypes        Vegan   Vegetarian
##          371          123          106
```

diet has 3 levels, and the other variables are binary with yes/no:

assemble_meals (binary) live_with_others (binary) gluten_free (binary)

First, I declare each of these four variables as a 'factor' variable

Then, JUST FOR THE BINARY VARIALBES, I create a numeric version of each variable in which yes = 1 and no = 0. As you'll see below, I do not do this coding for the one categorical variable with 3 levels.

I need to do this coding for the categorical variables because some R packages require such variables to be recognized as factors and some require that they are numeric.

```
mealdelivery.df$assemble_meals <- factor(mealdelivery.df$assemble_meals, levels=c("yes","no"), labels=c
mealdelivery.df$num_assemble_meals <- factor(mealdelivery.df$assemble_meals, levels=c("yes","no"), label
mealdelivery.df$num_assemble_meals <- as.numeric(as.character(mealdelivery.df$num_assemble_meals))

mealdelivery.df$live_with_others <- factor(mealdelivery.df$live_with_others, levels=c("yes","no"), label
mealdelivery.df$num_live_with_others <- factor(mealdelivery.df$live_with_others, levels=c("yes","no"), l
mealdelivery.df$num_live_with_others <- as.numeric(as.character(mealdelivery.df$num_live_with_others))

mealdelivery.df$gluten_free <- factor(mealdelivery.df$gluten_free, levels=c("yes","no"), labels=c("yes"
mealdelivery.df$num_gluten_free <- factor(mealdelivery.df$gluten_free, levels=c("yes","no"), labels=c(1
mealdelivery.df$num_gluten_free <- as.numeric(as.character(mealdelivery.df$num_gluten_free))

mealdelivery.df$diet <- factor(mealdelivery.df$diet, levels=c("AllFoodTypes","Vegan", "Vegetarian"), la
```

Research Question 1

What is the estimated mean satisfaction level according to a customer's diet type? Are there an important difference in mean satisfaction between these groups?

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

Diet is categorical (specifically, 3 levels), and satisfaction is numeric, so I use ANOVA to compare mean satisfaction levels between the three groups of customers.

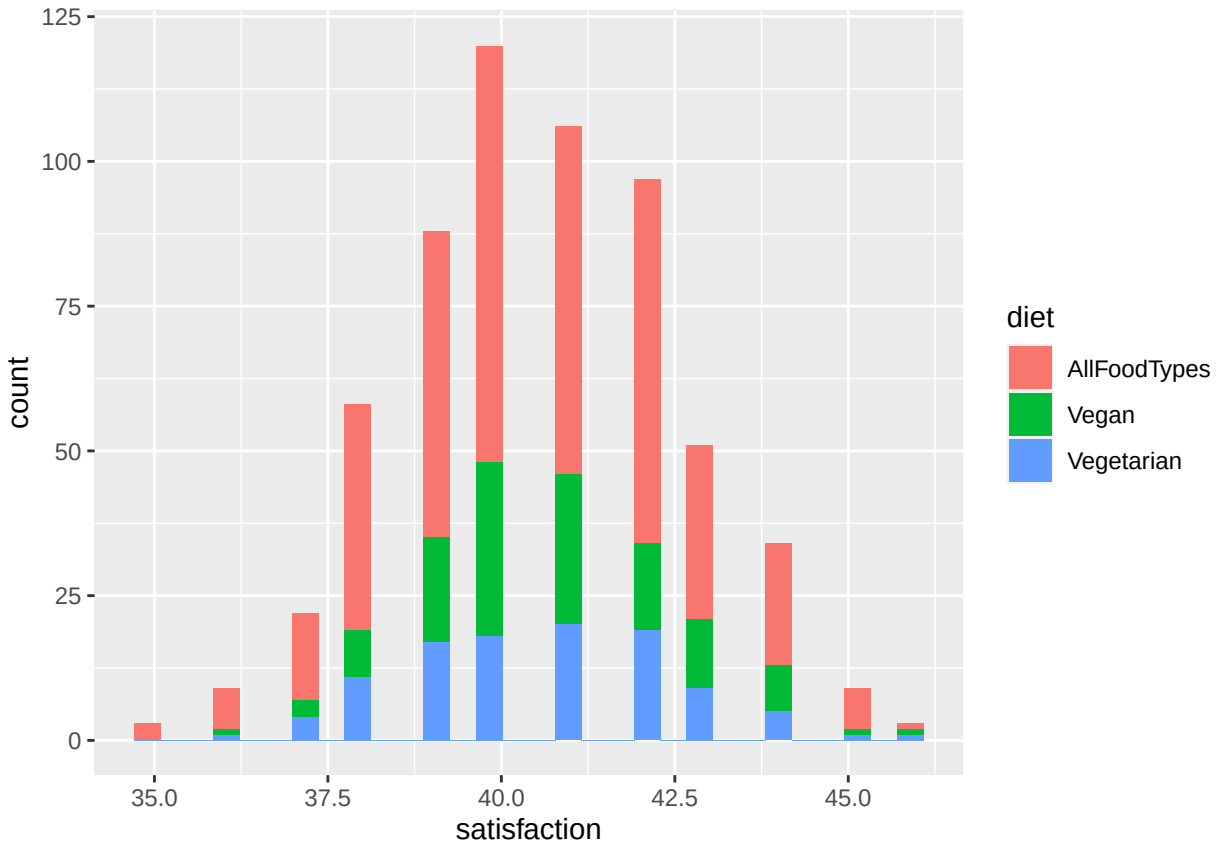
Null hypothesis: mean satisfaction is equal between the three diet groups (diet type and satisfaction are not related)

Alternative hypothesis: mean satisfaction differs at least between one group and the others (diet type and satisfaction are related)

First, I create a histogram that displays the satisfaction according to each of the three diet types.

```
ggplot(mealdelivery.df, aes(x = satisfaction, fill = diet )) + geom_histogram()
```

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



Now, I perform the ANOVA

```
aov_oneway =aov_ez(id = "cust.id",
       dv = "satisfaction",
       between = "diet",
       data = mealdelivery.df)
```

```
## Contrasts set to contr.sum for the following variables: diet
```

```
summary(aov_oneway)
```

```
## Anova Table (Type 3 tests)
##
## Response: satisfaction
##      num Df den Df   MSE      F       ges Pr(>F)
## diet      2    597 3.986 0.5951 0.0019898 0.5518
```

```
aov_oneway
```

```
## Anova Table (Type 3 tests)
##
## Response: satisfaction
##   Effect     df  MSE    F  ges p.value
## 1   diet 2, 597 3.99 0.60 .002    .552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1
```

I see that the p-value for the F statistic is greater than a significance level of .05, which indicates that the test result is not significant.

Thus, I fail to reject the null hypothesis.

The test result suggests that there are no differences in mean satisfaction between the three diet groups.

Before I deliver an executive summary to my client, I check the sample size of each of the diet groups.

```
table(mealdelivery.df$diet)
```

```
##
## AllFoodTypes        Vegan   Vegetarian
##          371          123          106
```

The sample size of each diet groups is greater than 30, so I don't need to worry about testing the assumption of normality. (If any one of the sample sizes had been <30, then I would need to test the assumption of normality using the Shapiro-Wilk test that I performed when checking this assumption for a t-test)

Next, I need to test the assumption of homogeneity of variance, and I use Levene's test for this.

The null hypothesis is that the variances are equal in the respective populations

The alternative hypothesis is that at least one variance in the respective populations differs from the others

```
leveneTest(satisfaction ~ diet, data = mealdelivery.df)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  1.4183 0.2429
##       597
```

The test statistic for Levene's test is not significant (the p value is $> .05$), so I do not reject the null hypothesis. I move forward in interpreting the test result from the ANOVA to my client.

Executive Summary: The three consumer groups based on the three diet types do not differ on average in their satisfaction with the product. Thus, any effort to improve satisfaction not need take into account the diet preferences of the consumers.

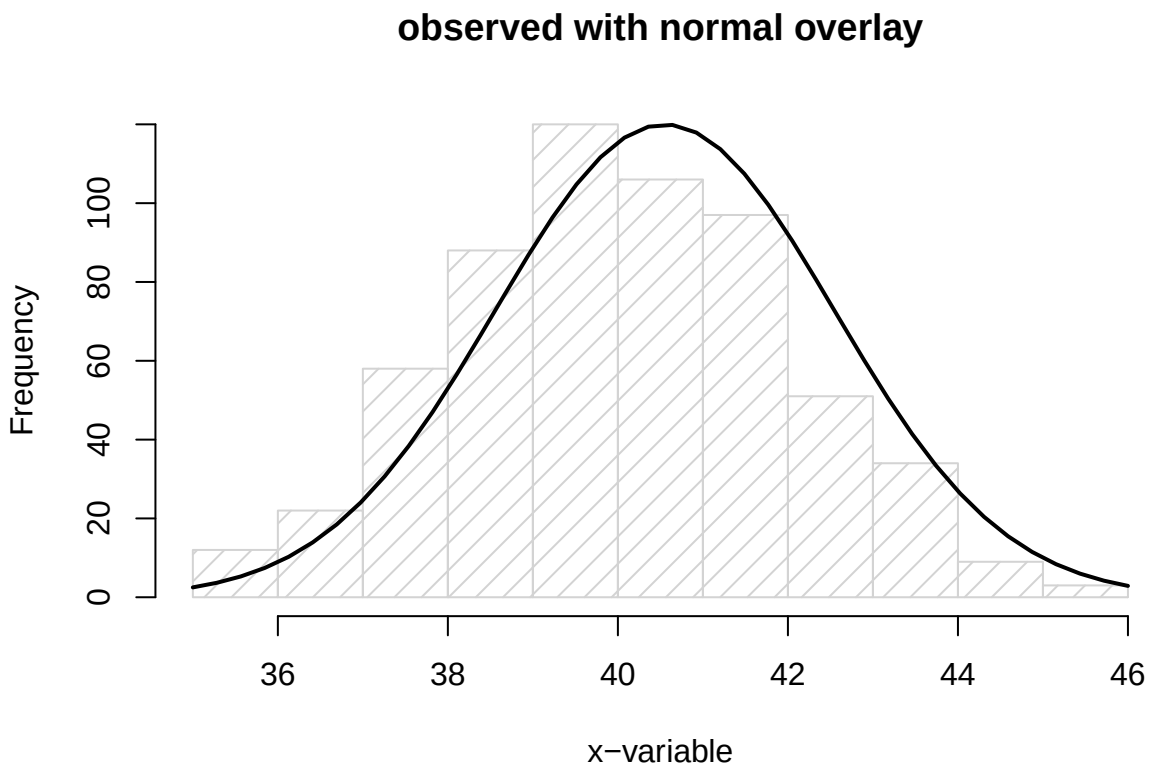
Research Question 2

The client wants to understand what variables are related to consumer satisfaction. Which variables are related to satisfaction and which are not?

The outcome variable is satisfaction. I examine the distribution of satisfaction scores to see if there are any outliers:

```
g = mealdelivery.df$satisfaction
h <- hist(g, breaks = 10, density = 10,
          col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")
xfit <- seq(min(g), max(g), length = 40)
yfit <- dnorm(xfit, mean = mean(g), sd = sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)

lines(xfit, yfit, col = "black", lwd = 2)
```



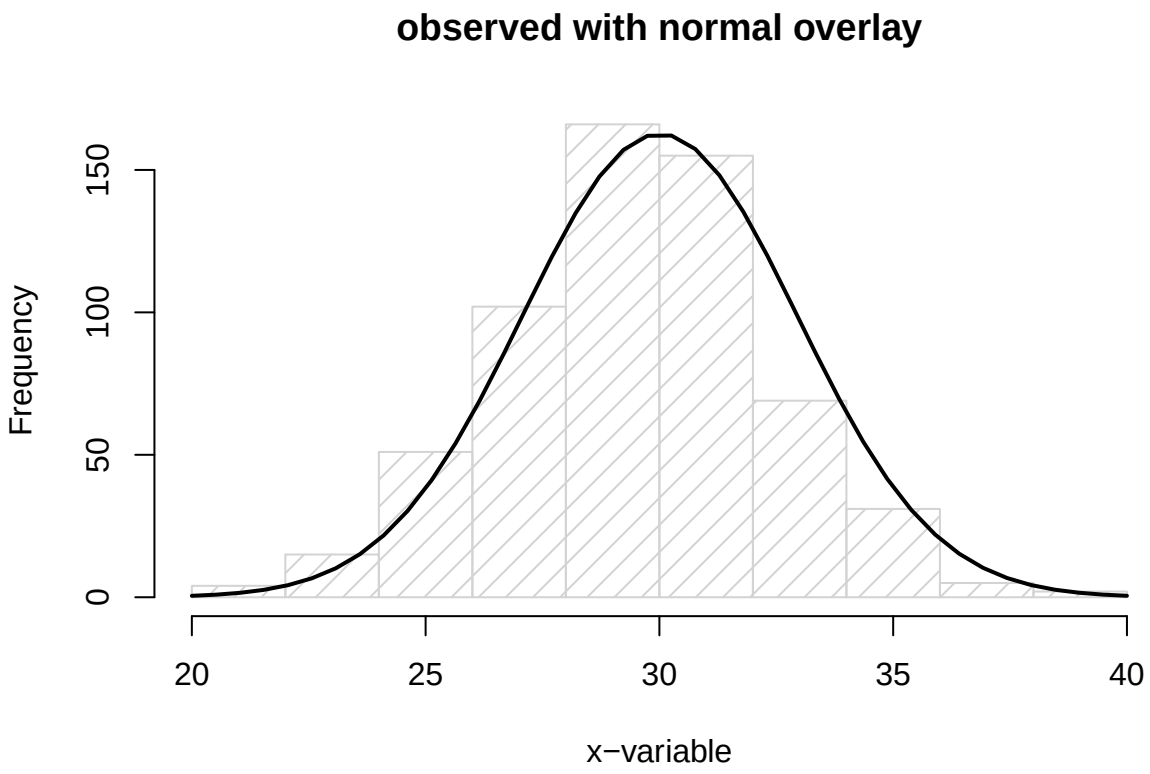
The distribution for Satisfaction looks fairly symmetric and I don't see any unusual values in either tail.

Next, let's take a look at the distributions of the numeric variables that I will test to see if they are related to satisfaction. Let's display each with a histogram and then in a scatterplot with satisfaction.

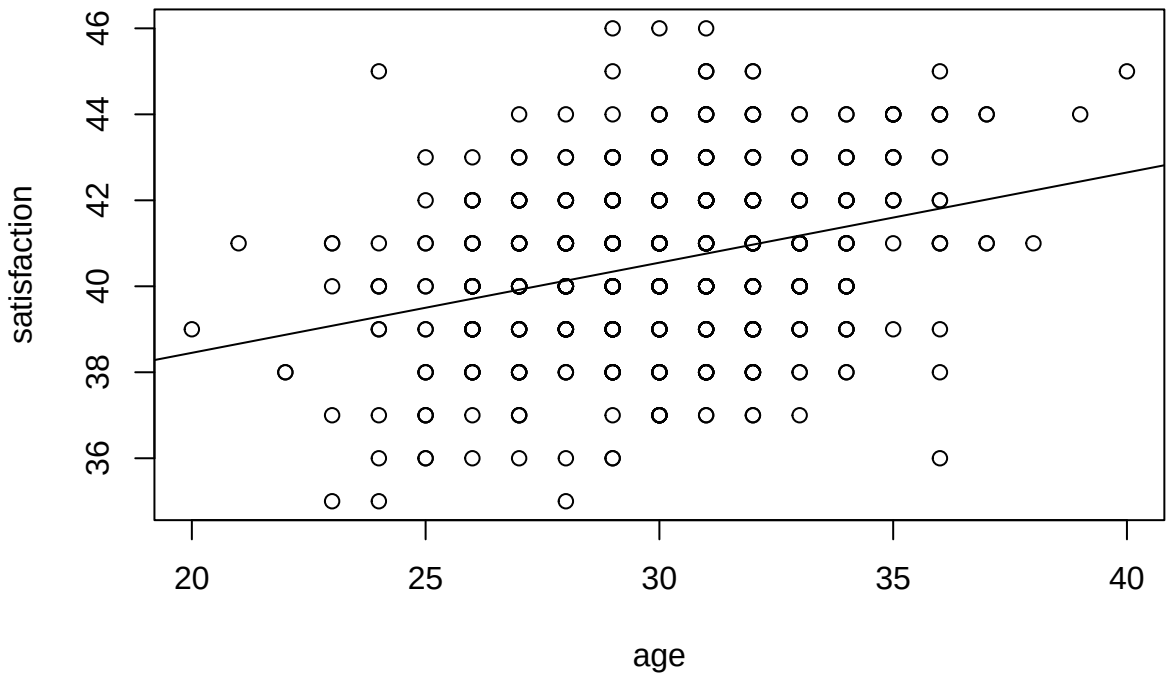
age

```
g = mealdelivery.df$age
h <- hist(g, breaks = 10, density = 10,
          col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")
xfit <- seq(min(g), max(g), length = 40)
yfit <- dnorm(xfit, mean = mean(g), sd = sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)

lines(xfit, yfit, col = "black", lwd = 2)
```



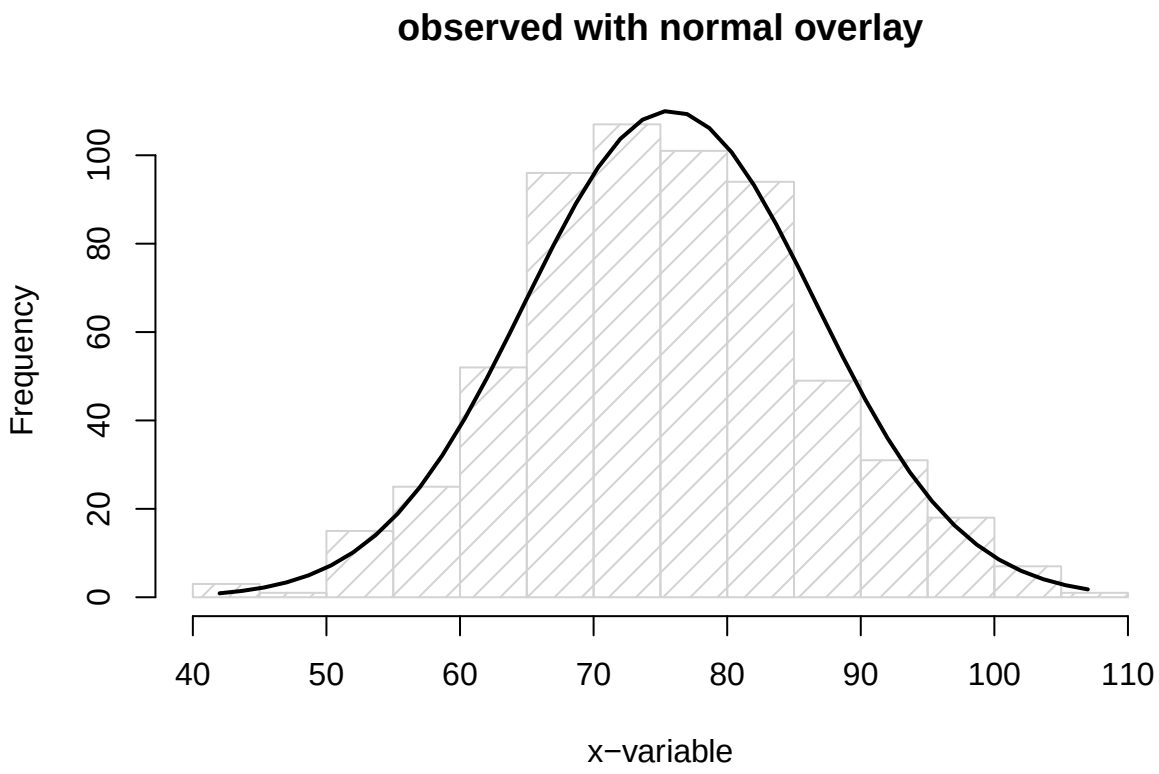

```
age_model <- lm(satisfaction ~ age,data=mealdelivery.df)
plot(satisfaction ~ age, data=mealdelivery.df)
abline(age_model)
```



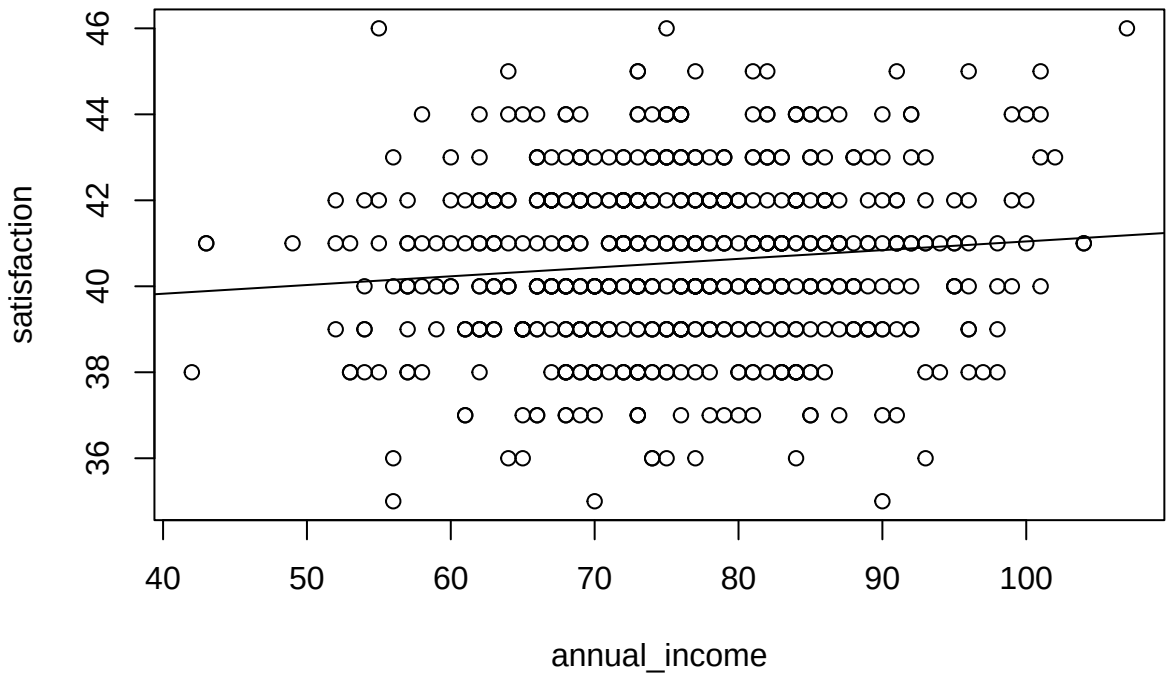
annual_income

```
g = mealdelivery.df$annual_income
h <- hist(g, breaks = 10, density = 10,
          col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")
xfit <- seq(min(g), max(g), length = 40)
yfit <- dnorm(xfit, mean = mean(g), sd = sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)

lines(xfit, yfit, col = "black", lwd = 2)
```



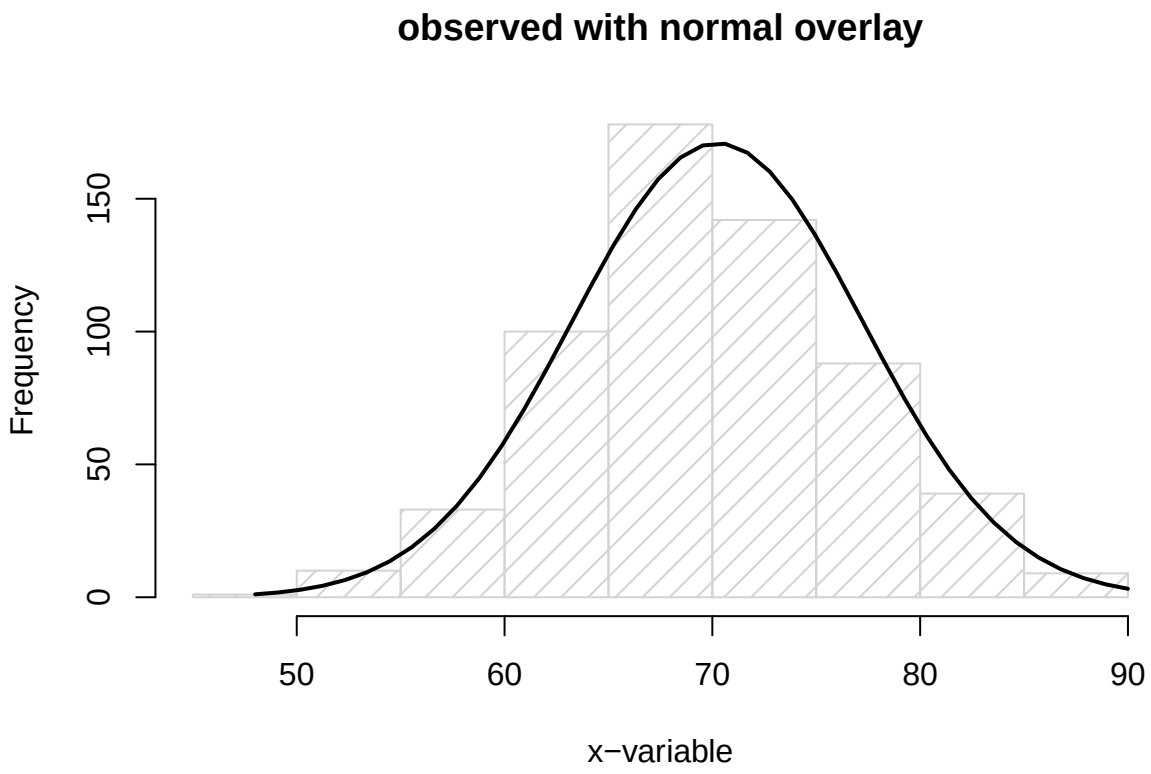

```
annual_income_model <- lm(satisfaction ~ annual_income,data=mealdelivery.df)
plot(satisfaction ~ annual_income, data=mealdelivery.df)
abline(annual_income_model)
```



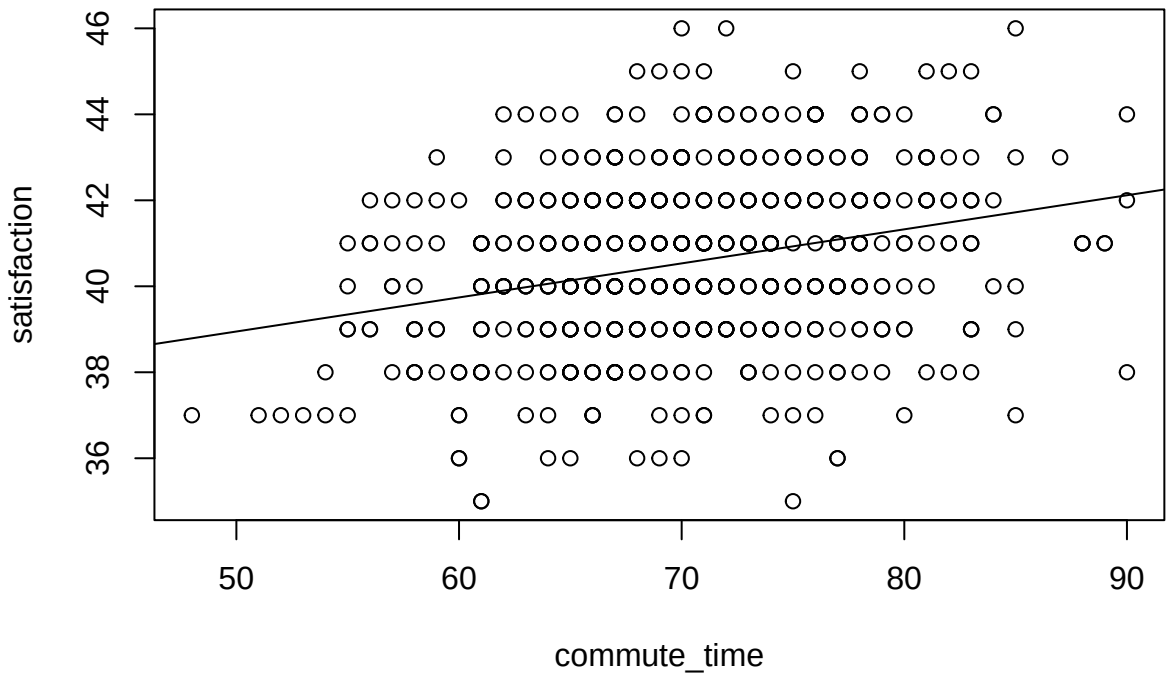
commute_time

```
g = mealdelivery.df$commute_time
h <- hist(g, breaks = 10, density = 10,
          col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")
xfit <- seq(min(g), max(g), length = 40)
yfit <- dnorm(xfit, mean = mean(g), sd = sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)

lines(xfit, yfit, col = "black", lwd = 2)
```



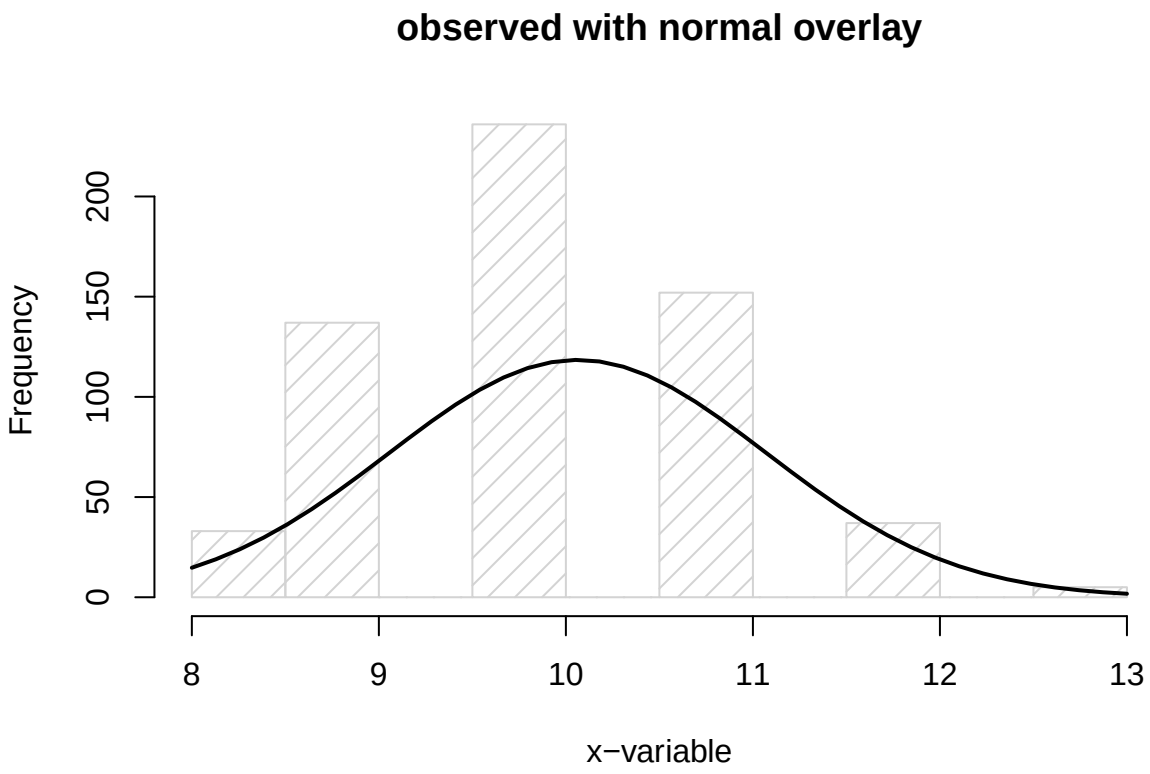

```
commute_time_model <- lm(satisfaction ~ commute_time,data=mealdelivery.df)
plot(satisfaction ~ commute_time, data=mealdelivery.df)
abline(commute_time_model)
```



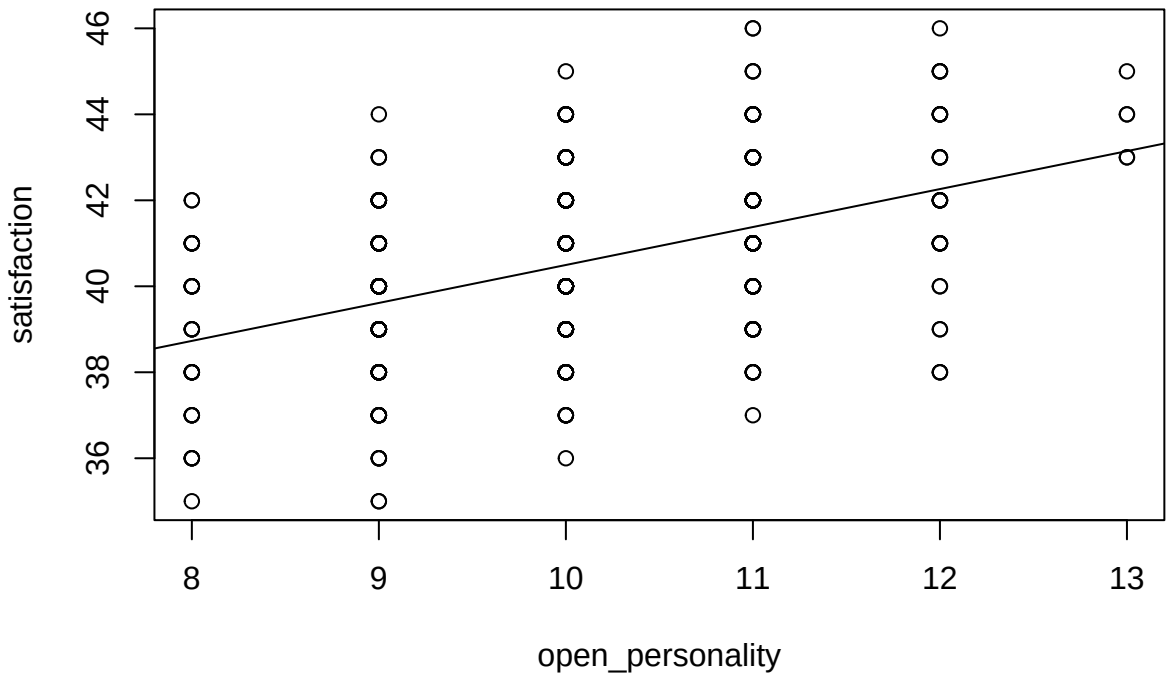
open_personality

```
g = mealdelivery.df$open_personality
h <- hist(g, breaks = 10, density = 10,
          col = "lightgray", xlab = "x-variable", main = "observed with normal overlay")
xfit <- seq(min(g), max(g), length = 40)
yfit <- dnorm(xfit, mean = mean(g), sd = sd(g))
yfit <- yfit * diff(h$mids[1:2]) * length(g)

lines(xfit, yfit, col = "black", lwd = 2)
```




```
open_personality_model <- lm(satisfaction ~ open_personality,data=mealdelivery.df)
plot(satisfaction ~ open_personality, data=mealdelivery.df)
abline(open_personality_model)
```



I use multiple regression to test which variables are related to satisfaction with the product. I choose multiple regression because I have one numeric outcome variable and a set of explanatory variables (some of which are categorical and some are numeric)

The analysis provides a test for each explanatory variable. There is a null and an alternative hypothesis for the effect of each explanatory variable on the outcome variable:

Null hypothesis: The variable is not related to satisfaction

Alternative hypothesis: The variable is related to satisfaction

Let's first examine the correlations between each explanatory variable and satisfaction, as well as the correlations between the explanatory variables. Only numeric variables can be used here, so this is where I need to use the numeric values for the binary variables and I won't include 'diet' because it is categorical with 3 levels.

Using the corrplot package, I need to specify the columns of each variable that I want to include in the display. Let's check the order of the variables in the data frame:

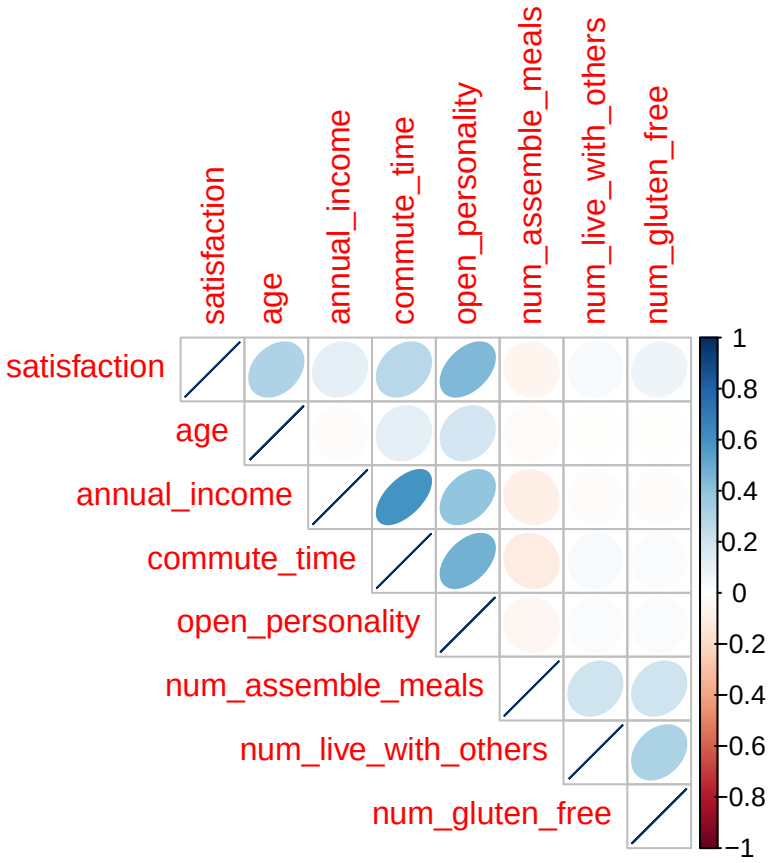
```
str(mealdelivery.df)
```

```
## 'data.frame':    600 obs. of  13 variables:
##  $ cust.id             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ assemble_meals      : Factor w/ 2 levels "yes","no": 1 2 2 2 2 2 2 2 1 2 ...
##  $ live_with_others    : Factor w/ 2 levels "yes","no": 2 1 2 1 1 1 1 1 1 1 ...
##  $ gluten_free         : Factor w/ 2 levels "yes","no": 2 2 2 2 1 1 1 2 2 2 ...
##  $ diet                : Factor w/ 3 levels "AllFoodTypes",..: 1 2 1 1 2 2 2 2 2 1 ...
##  $ satisfaction        : int  39 40 41 38 42 43 39 40 41 39 ...
```

```
##  $ age                 : int  26 26 31 31 28 35 29 32 34 32 ...
##  $ annual_income       : int  76 67 82 94 67 67 92 63 52 77 ...
##  $ commute_time        : int  71 69 76 82 63 59 68 65 66 73 ...
##  $ open_personality    : int  9 11 11 11 9 11 11 9 10 9 ...
##  $ num_assemble_meals  : num  1 0 0 0 0 0 0 0 1 0 ...
##  $ num_live_with_others: num  0 1 0 1 1 1 1 1 1 1 ...
##  $ num_gluten_free     : num  0 0 0 0 1 1 1 0 0 0 ...
```

In the corrplot package, I'd like have satisfaction (the 6th variable) to appear first in the display, followed by the explanatory variables (excluding diet) and using the numeric versions of the three binary variables:

```
corrplot(cor(mealdelivery.df[ , c(6,7:13)]),method='ellipse',type = 'upper')
```



Now, I perform the multiple regression.

Using corrplot, I could not use categorical variables. That is, the package allows only numeric variables.

The multiple regression model will allow for categorical variables, so I can include diet in this analysis.

```
model1 <- lm(satisfaction ~ age + annual_income + commute_time + open_personality + num_assemble_meals +

summary(model1)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + annual_income + commute_time +
##     open_personality + num_assemble_meals + num_live_with_others +
```

```
##     num_gluten_free + diet, data = mealdelivery.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.635 -1.197  0.020  1.184  4.519
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          27.445518   1.021674  26.863  < 2e-16 ***
## age                   0.148001   0.024508   6.039 2.75e-09 ***
## annual_income        -0.022278   0.008201  -2.717  0.00679 **
## commute_time          0.037610   0.013268   2.835  0.00475 **
## open_personality      0.764385   0.081541   9.374  < 2e-16 ***
## num_assemble_meals   -0.196079   0.153267  -1.279  0.20128
## num_live_with_others -0.001820   0.170945  -0.011  0.99151
## num_gluten_free       0.322512   0.198226   1.627  0.10427
## dietVegan             0.100523   0.201839   0.498  0.61865
## dietVegetarian        0.034063   0.220966   0.154  0.87754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.714 on 590 degrees of freedom
## Multiple R-squared:  0.2727, Adjusted R-squared:  0.2616
## F-statistic: 24.58 on 9 and 590 DF,  p-value: < 2.2e-16
```

The p-values for the t statistics for age, annual_income, commute_time, and open_personality are less than a significance level of .05.

So, for these 4 explanatory variables, I reject the null hypothesis and conclude that each relates to satisfaction (taking into account the associations that each variable in the model has with satisfactions; that is, the individual estimated effects take into account the associations that other variables have with the outcome).

The p-values for the t statistics for num_assemble_meals, num_live_with_others, diet and num_gluten_free are greater than a significance level of .05.

So, for these 4 explanatory variables, I fail to reject the null hypothesis and conclude that none of these variables relate to satisfaction (taking into account the associations that each variable in the model has with satisfactions).

By examining the Adjusted R-squared value (adjusted R squared is less biased relative to the Multiple R squared, so the Adjusted R squared value is usually reported), I see that as a set, the explanatory variables account for about 26% of the variation in satisfaction. This indicates that the client may want to invest more effort into finding other variables that are driving satisfaction.

Let's now see if I can simplify the model by dropping the 4 variables that were not statistically significant. I'll compare the fit of the reduced model (I'll call this reduced model 'Model 2') to the full model.

```
model2 <- lm(satisfaction ~ age + annual_income + commute_time + open_personality, data=mealdelivery.df)

summary(model2)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + annual_income + commute_time +
##     open_personality, data = mealdelivery.df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6528 -1.2161 -0.0014  1.1719  4.7240
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      27.294866   1.008398  27.068  < 2e-16 ***
## age               0.148148   0.024463   6.056 2.47e-09 ***
## annual_income    -0.022453   0.008175  -2.747  0.00620 **
## commute_time      0.039723   0.013198   3.010  0.00273 **
## open_personality  0.767282   0.081122   9.458  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.714 on 595 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.2617
## F-statistic: 54.09 on 4 and 595 DF,  p-value: < 2.2e-16
```

```
anova(model2,model1)
```

```
## Analysis of Variance Table
##
## Model 1: satisfaction ~ age + annual_income + commute_time + open_personality
## Model 2: satisfaction ~ age + annual_income + commute_time + open_personality +
##     num_assemble_meals + num_live_with_others + num_gluten_free +
##     diet
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    595 1748.6
## 2    590 1734.2  5    14.386 0.9789 0.4299
```

From the ANOVA test used to compare the fit of these two models, the test statistic has a p value that is greater than .05, so the fit of the full model is not statistically different from the fit of the reduced model. In fact, the adjusted R squared value is still at about 26%.

Executive Summary: Satisfaction with the product is not related to diet type, whether the meal needs assembling, whether the customer lives with others, or whether they have a gluten-free diet. Specifically, age, commute_time and open_personality are positively related to satisfaction, and annual_income is negatively related to satisfaction. Older customers, those who have a relative long commute time and are relatively more open to new experiences are relatively more satisfied with the product. Those with a relatively low annual income are relatively less satisfied with the product.