Data Analysis Report

Pets ownership and Happy &Healthy Degree

Yufei Wang

Ting Gong

Introduction

Social support is common and important for psychological and physical well-being. Under high pressures from living and working, more people are going to ask therapists for help. But are people the only source for our sense of belongingness and the support of our physical health? In reality, more people raise dogs or cats to quell their stress and loneliness, to boosting the productivity and the overall well-being. Pets are believed to play a significant role in people`s daily lives. Research conducted in the Journal of Personality and Social Psychology indicates that pets confer significant benefits to their owners and improve their lives. For instance, introverted people are more likely to enjoy positive consequences from their pets, owning a cat or dog can increase chances of surviving a heart attack, and pet owners who are facing serious health challenges are less fearful to medical treatments. As you can see, the consequences for people owning pets are impressing and amazing, and many researches are conducted on this topic with different observations.

In this observational study, we focused on in the mean difference of happy degree and healthy degree with those who own pets and those who do not own pets. The sampling strategy follows these steps: (1)Select that the target population is the overall population at Washington State; (2) Select that the accessible population is the population in South Lake Union, Seattle, WA; (3) Outline the sampling collection plan; (4) Use a simple random sampling method and all possible samples of objects are equally likely to occur (5)Recruit the samples in South Lake Union area. We designed a survey (in Appendix) with total nine questions. Measurements are following: Categorical variables are Gender, Age, Occupation, Pet Type, Happy score and Healthy score .Continuous variables are Walk Pet time and Monthly money spent on pets. We randomly selected 55 people in SLU area, and handed the survey to them. In the exploratory analysis, we will provide detailed graphs to help readers visualize and understand the important aspects of the dataset.

There are two inevitable biases in this survey and collection method: sampling bias and response bias. Sampling bias is introduced when some members of the intended population are less likely to be surveyed than others, and it can cause an overestimate or underestimate of the observations. In this observational study ,we only sent the survey directly to people who are in SLU in regular workdays .The way to remedy this trouble is to share the survey via various methods such as email and website. Response bias comes from less-than-truthful responses. For example, people want to be agreeable, and tell you what you want to hear. In this observational study, we used the number (1 to 10) to define the degree of happiness and the degree of healthiness, and the sensitivity of these questions may contribute to the response bias. The way to reduce it is to ask more neutrally and generally worded questions and to make sure that the answer options are not leading.

Exploratory analysis and data visualization

1. **The overall sample distribution**

Given the objectives of our study, we divided respondents into two groups based on whether they own pets or not. In this study, we collected surveys from 55 respondents in South Lake Union. Figure 1 shows that 69% of respondents own pets, and 31% of respondents do not own pets. The reason that the number of people who own pets exceeds the number of people who do not own pets are following:

1) Population of pet owner in South Lake Union is very large.

2) We did not strictly follow random sampling rules to collect, which could lead to the sampling bias.

Next, we counted the numbers of females and males , and used a bar plot to describe the gender distribution of two sample groups, which helped us familiarize with our data.

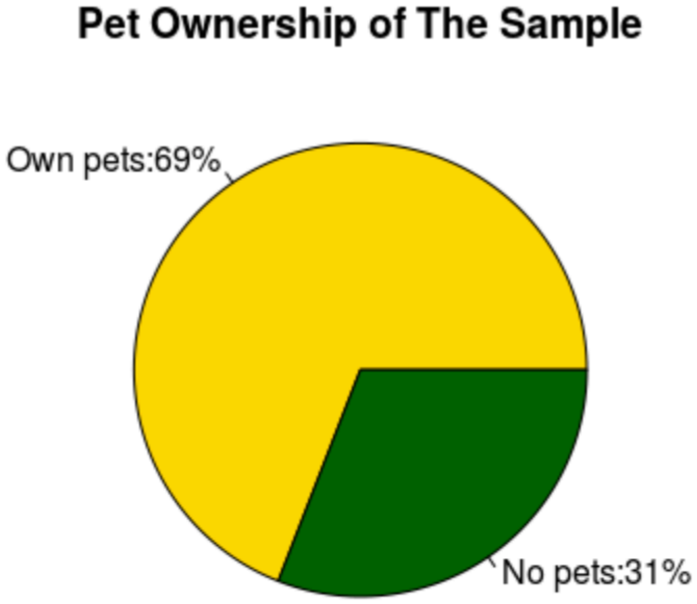


Figure 1 Pet ownership of the sample

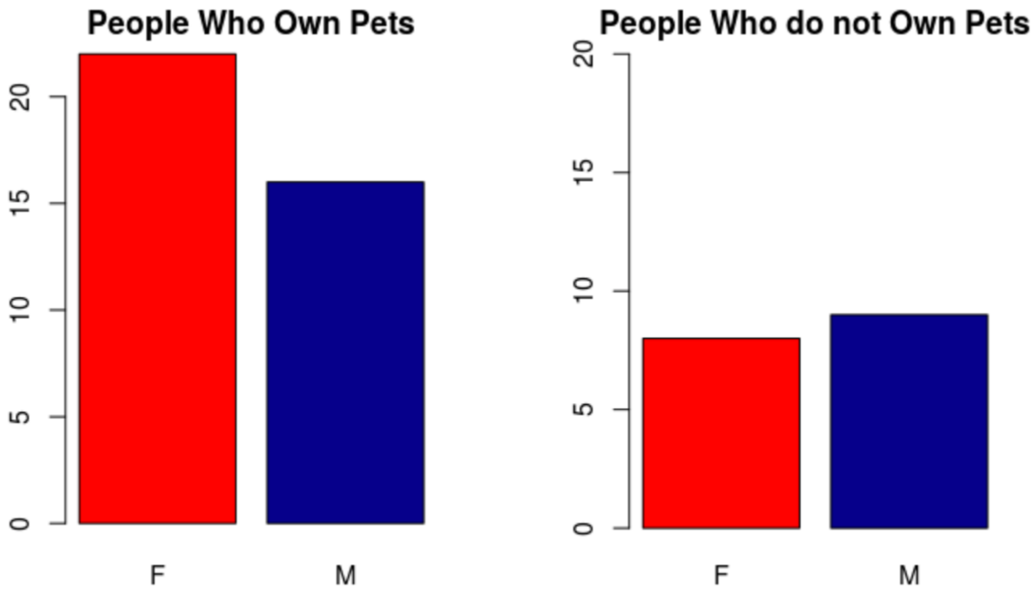


Figure 2 The bar plot of Gender distribution of two sample groups

2. **The sample distribution of Happy Score vs Two Sample Groups**

In order to understand sample groups better, we used box plots to compare the distribution of the two samples of happy scores (Figure 3). The ranges of happy scores are all from 5 to 10. Score 10 and score 5 are outliers in the data of pet owners. The quartiles and Whiskers of both groups are same. Since the median for both groups are closer to lower quartile, we can assume that two distributions are skewed to the left. There are no obvious differences between 2 box plots.

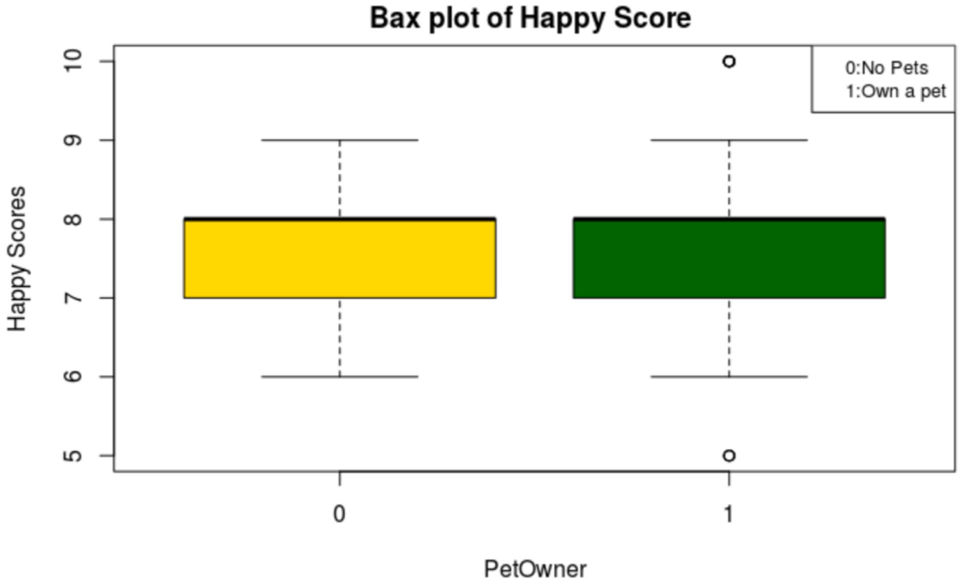


Figure 3 Box plot of Happy Score vs Two Sample Groups

3. The sample distribution of Healthy Score vs Two Sample Groups

In order to understand sample groups better, we used box plots to compare the distribution of the two samples of healthy scores (Figure 4). The graph displays that the distribution of people who do not own pets is left skewed, and the distribution of pet owners is right skewed. The health score of the sample population ranges from 4 to 10. There are two outliers (score 10 and score 5)in the data of people who do not own pets, and three outliers (score 10, score 5 ,score 4) in the data of pet owners. There are no obvious differences between 2 box plots.

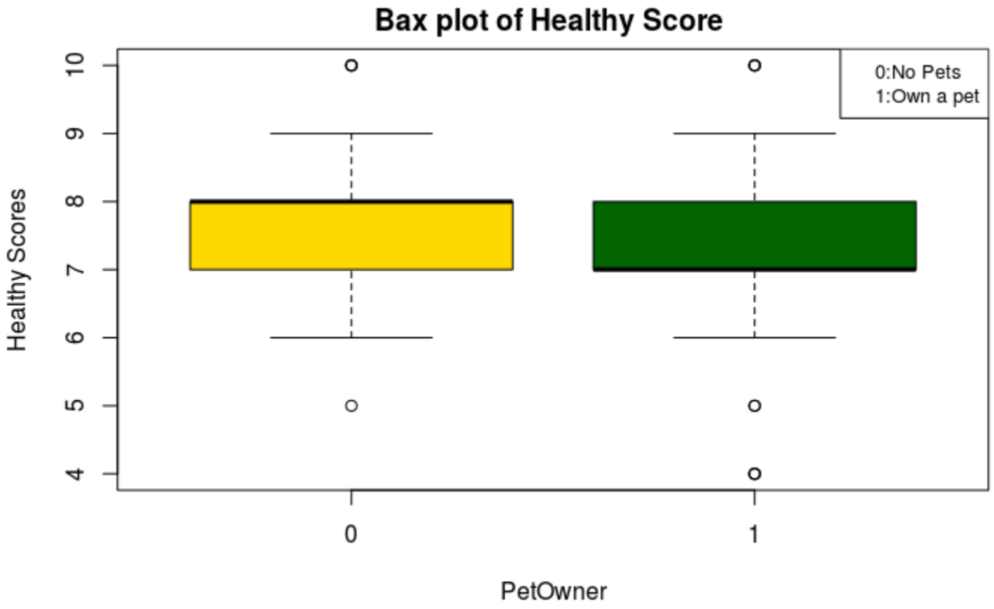


Figure 4 Box plot of Healthy Score vs Two Sample Groups

In summary, the happy score of people who own pets ranges from 5 to10, and the happy score of people who do not own pets ranges from 6 to 9, which indicate that most respondents are satisfied with their current lives. More than 80% of respondents` healthy scores are above 7, which shows that most respondents are in good health conditions.

4. **The sample distribution of Age vs Scores**

Among two sample groups, lines of best fit on four scatter plots show the negative trend between age and scores. Data points are not actually clustered along the line, therefore, we concluded that there are no relationships between ages and scores.

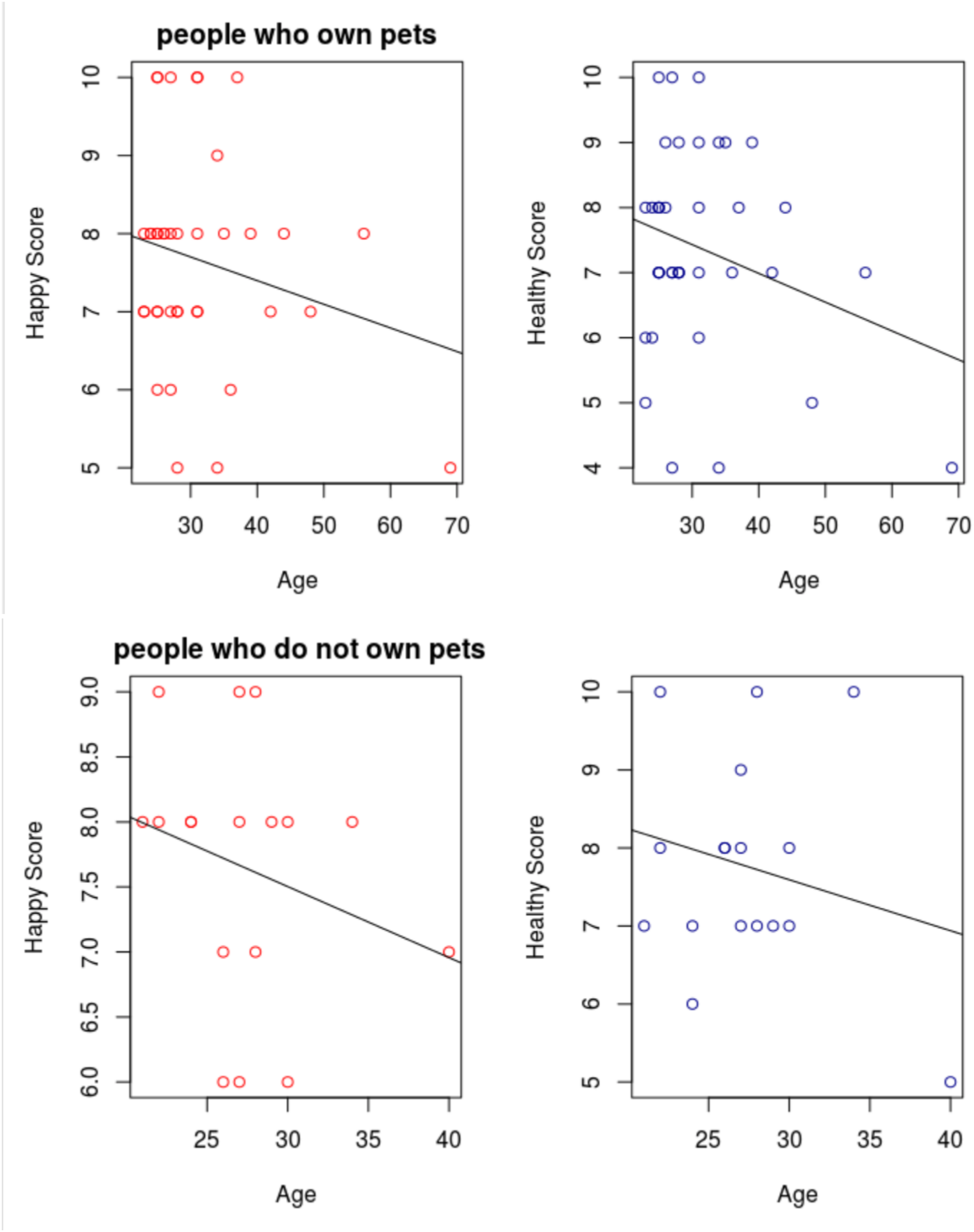


Figure 5 The scatter plot of Age vs Scores

5. The sample distribution of Pet Time Vs Scores

The time for walking pets ranges between 10 minutes to 3 hours within different happy scores. Information we interpreted from the box plot of happy scores and pet time are following: Except for score 9 (because of the limited sample size), the median time of other happy scores is 1 hour. Given the much longer "whiskers" for score 7, 8 and 10, we could interpret that pet time can vary widely for the same happy score.

Figure 6 is the box plot of healthy score and pet time, and it shows that except for score 5 and 6 (because of the limited sample size), the median pet time of other healthy scores is 1 hour. The interquartile ranges of healthy score 9 and 10 are similar, though the overall range of the score 9 data set is greater for the score 10. The length of the box for score 9 is more than twice that of the other boxes when scores equal from 4 to 7, and we could interpret that pet time can vary widely for the same healthy score, especially for the higher healthy score.

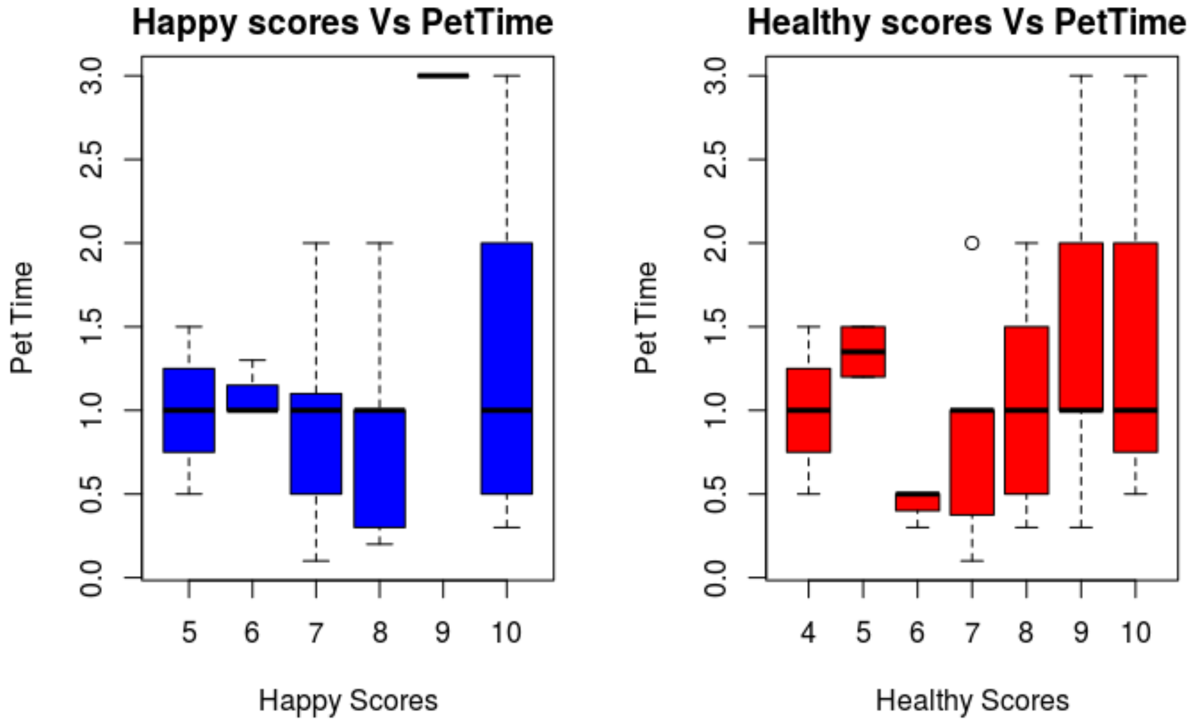


Figure 6 The box plot of Pet Time vs Scores

6. The sample distribution of Monthly Money Vs Scores

The monthly money ranges from 15 to 1000 dollars in this survey. From Figure 7, the lengths of each box are comparatively short, and most of respondents spend less than $200 a month on their pets. The median of monthly month of higher scores is higher than the median of low scores, therefore, we could indicate that the healthy scores are positively related to the money that a person spends monthly on pets.

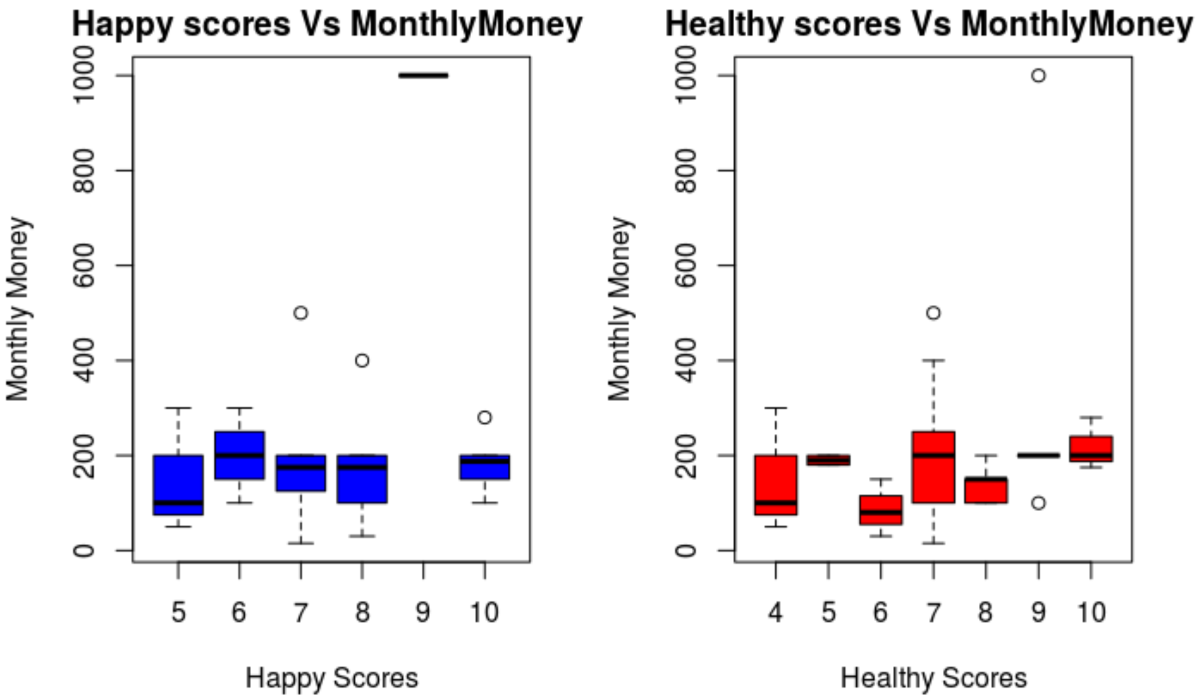


Figure 7 The box plot of Monthly Money vs Scores

Statistical Analysis

Two Sample Tests

1. Questions of interest are with the difference in means between two populations: people who own pets and people who do not own pets
2. Assumptions to use the t-test: 1)The observations are independent. 2) Sample size is larger than 30. When the sample size is larger than 30, then the sample variance is a very good estimate for population variance and the t distribution will be very close to normal distribution N(0,1).
3. Observations belong to different populations. It is possible to combine all observations into a single group and do a single sample test, but that will result in the answer to a different question.

Questions of interest

In our survey, we wanted to answer the following questions :

Do people who own pets have the same happy degree as people who do not own pets?

Do people who own pets have the same healthy degree as people who do not own pets?

We are interested in the difference in population means and we have sample mean and variance. Therefore, we can use a t-test for the difference in means

Check the t-test condition

The size of representation samples is 55 and the sample size of people who own pets is 38, therefore, both can be assumed normally distributed. Since the sample size of people who do not own pets is less than 30, we need to confirm that this sample seems to be normally distributed. We used QQ plot to check this condition, and showed results as following:

QQ plot of standard normal distribution

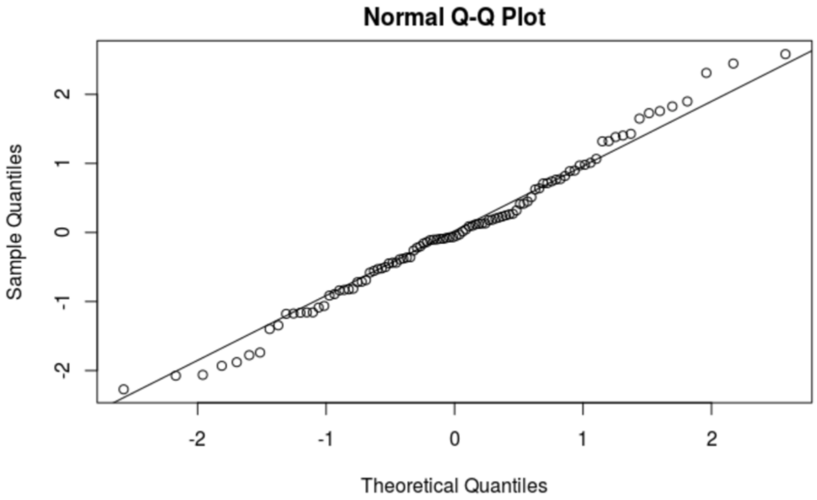


Figure 8 Normal QQ plot

This plot indicates that the points in the QQ-normal plot lie on a straight diagonal line when it is the normal distribution .We add the line to this QQ plot and the deviations from the straight line are minimal.

Happy Score:

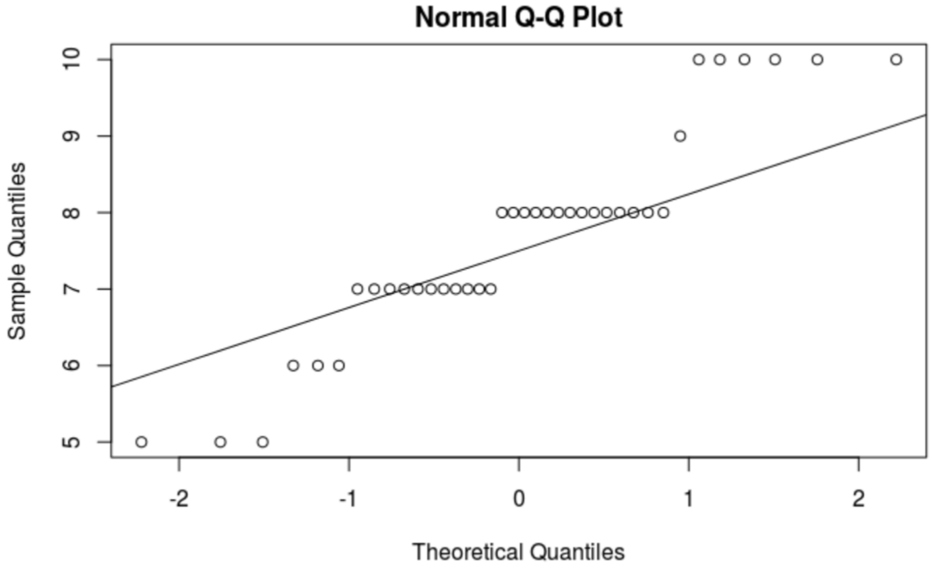


Figure 9 Normal QQ plot of sample of people who own pets

Healthy Score:

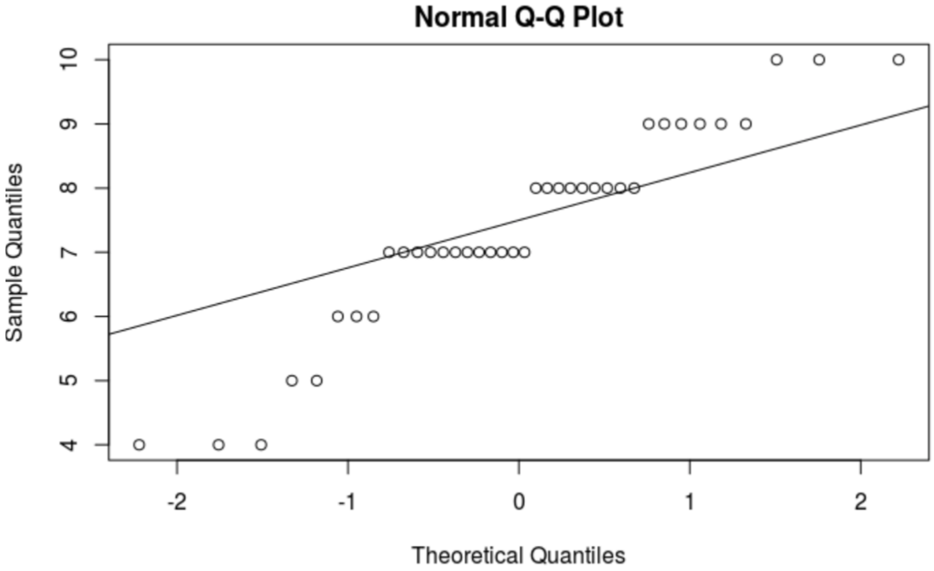


Figure 10 Normal QQ plot of sample of people who own pets

Happy Score:

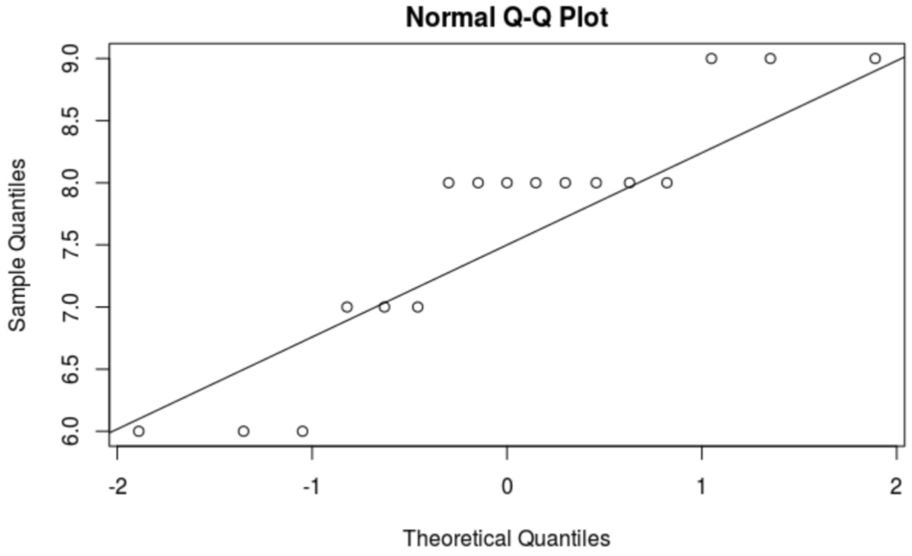


Figure 11 Normal QQ plot of sample of people who do not own pets

Healthy Score:

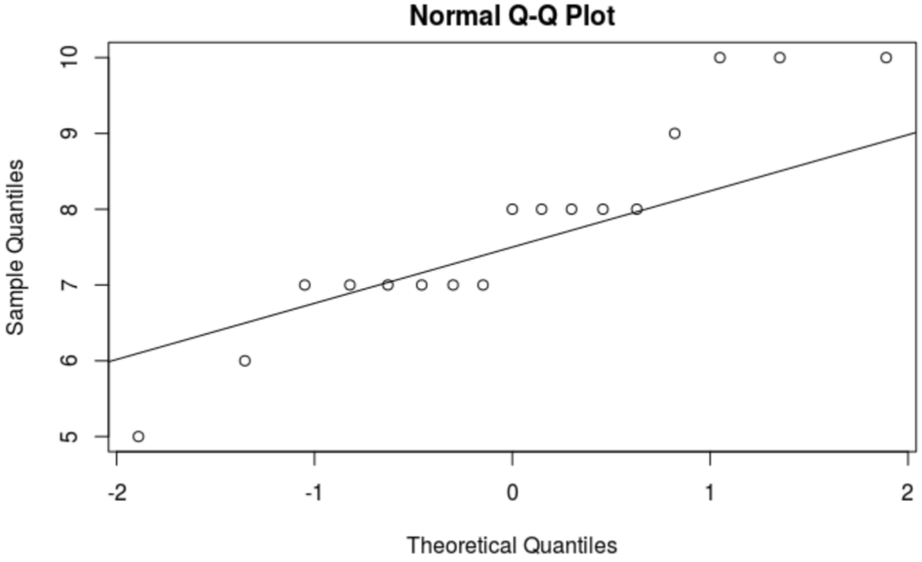


Figure 12 Normal QQ plot of sample of people who do not own pets

The QQ plot doesn't look good when the sample size is less than 30. But we still proceed with the test and knew that the reliability of our test results and confidence interval are reduced. We will discuss detailed explanation for the limitation in discussion.

Parameter

We are interested in the true population mean difference in happy degree between those who own pets and those who do not own pets at South Lake Union : μ_{p1}-μ_{n1}

We are interested in the true population mean difference in healthy degree between those who own pets and those who do not own pets at South Lake Union: μ_{p2}-μ_{n2}

Hypothesis Test 1

- The null hypothesis, H_0: μ_{p1}-μ_{n1}=0

  The true population mean in happy degree for those who own pets is equal to those who do not own pets.
- The alternative hypothesis, H_1: μ_{p1}-μ_{n1} ≠0

  The true population mean in happy degree for those who own pets is different from those who do not own pets.

Hypothesis Test 2

- The null hypothesis, H_0: μ_{p2}-μ_{n2}=0

  The true population mean in healthy degree for those who own pets is equal to those who do not own pets.
- The alternative hypothesis, H_1: μ_{p2}-μ_{n2} ≠0

  The true population mean in healthy degree for those who own pets is different from those who do not own pets.

Choose a significance level: α=0.05

Sample Statistics

- Hypothesis Testing 1:$\bar{x}_{p1}$-$\bar{x}_{n1}$
- Hypothesis Testing 2: $\bar{x}_{p2}$-$\bar{x}_{n2}$

Test Statistic

$$t_{\min\left(n_{p-1,}n_{n-1,}\right)}=\frac{(\bar{x}_p-\bar{x}_n)-(\mu_p-\mu_n)}{\sqrt{\frac{s_p^2}{n_p}+\frac{s_n^2}{n_n}}}$$

P-value

We did two-sided based on the alternative hypothesis with 95% confidence interval and put detailed R codes for t-test of a knitted R markdown file in Appendix.

The results are following:

Hypothesis testing 1: P-value = 0.9769

Hypothesis testing 2: P-value = 0.3784

Confidence Interval

- Two-sided

$$\left(\bar{x}_p-\bar{x}_n\right)-t_{\min\left(n_{p-1,}n_{n-1,}\right),\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_p}+\frac{s_n^2}{n_n}}<\mu_p-\mu_n<\left(\bar{x}_p+\bar{x}_n\right)-t_{\min\left(n_{p-1,}n_{n-1,}\right),\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_p}+\frac{s_n^2}{n_n}}$$

From R, results of 95% confidence interval are following

Hypothesis testing 1: lower bound = -0. 7366; upper bound = 0.7523

Hypothesis testing 2: lower bound = -1.2911; upper bound = 0.4985

Interpretation

There is some evidence to suggest that we fail to reject the null hypothesis. P-value from two hypothesis tests are larger than 0.05,which indicates that the true population means in happy degree and healthy degree for those who own pets are same as those who do not own pets. With 95% confidence, the true difference in the mean in happy degree between those who own pets and those who do not is between -0.7366 and 0.7523. With 95% confidence, the true difference in the mean in healthy degree between those who own pets and those who do not is between -1.2911 and 0.4985. The null hypothesized difference between the mean in happy degree and healthy degree are zero, and zero is in the 95% confidence interval for both tests, which is consistent with the failure of the rejection of the null hypothesis.

Histogram of the sampling distributions

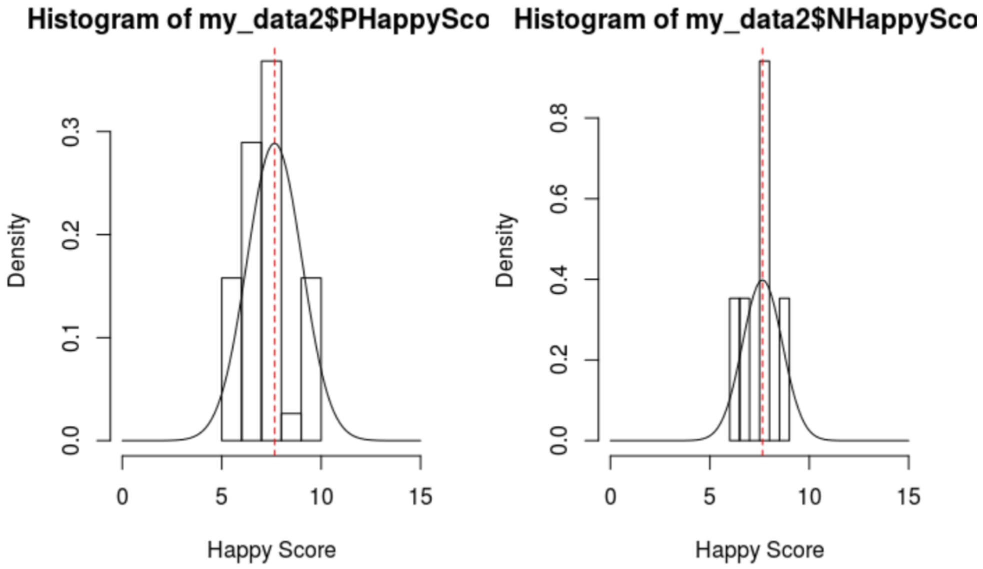


Figure 13 Histogram of two sample populations of Happy score

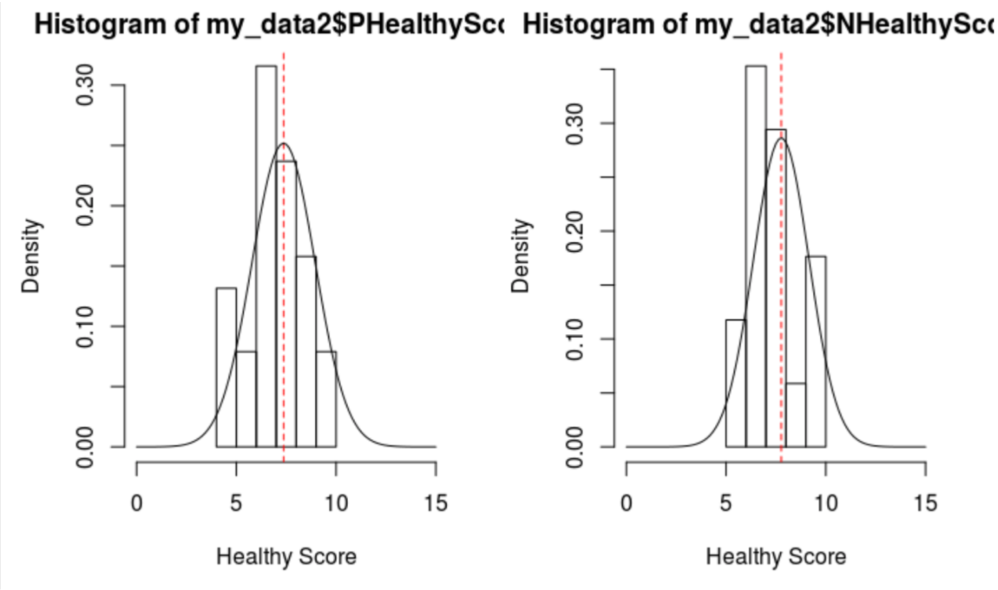


Figure 14 Histogram of two sample populations of Healthy score

Sampling distribution is the population distribution of a variable describes the values of the variable for all individuals in a population. In our study, we plotted the probability density of happy scores and healthy scores between those who own pets and those who do not own pets Red lines represent sample means. As you can see, the densest parts are all between 5 and 10 for two sampling populations. The curve represents the normal distribution with the same mean and the same variance in the plot of the original sampling distribution. To summarize, the distribution of sample means will be approximately normal as long as the sample size is large enough.

Discussion

- Summary of your findings.

  Most participants are quite satisfied with their current life, and consider themselves in good health. There are no significant differences in happy degree and healthy degree among the two sample groups based on the result of t-test. According to scatter plots, there are no relationships between ages and scores. Time spending on pets varies widely for high happy score and healthy score. The amount of money that a person spends monthly on pets on the highest score of happiness and health is generally greater than others. In addition, from the box plot of healthy score and monthly money, there could be a positive relationship when the healthy score is higher than 8.

- Implications of your findings.

As the introduction mentioned, pets have significant physical and psychological benefits to their owners. It is important to understand how pets can influence all aspects of our lives. Although, the conclusion of this study is that there are no differences in true population means in the happy degree and healthy degree between people who have pets and those who do not have pets. Due to sampling bias, respond bias, and confounding variables in this survey, we cannot completely neglect the positive effects of pets on human happiness and health. The sample data collected by this research, such as gender, age, time spent with pets and monthly expenses are also very useful for studying behaviors of pet owners. In addition, we can expand this study to conduct further research about many other positive roles that pets play in people's lives, for example, service dogs are trained to aid blind or deaf owners, diabetic alert dogs can recognize drops in blood sugar levels.

- Limitations and Extensions

1. We designed an observational study to study the probable cause and effect associated with the happy degree and healthy degree with those who own pets and those who do not own pets. In introduction, we discussed about the sampling bias and response bias in the design phase. However, those biases cannot be completely corrected after the completion of a study, thus we have to minimize their impact during the analysis phase.
2. Confounding bias is a systematic error in inference due to the influence of confounding variables. Confounding variables are extra variables that we do not account for in the study. Besides the pet factor, multiple factors can influence human happiness and health. For example, people who have strong support network may feel happier than people who are lonely because they feel a sense being a part of community. People have higher education level and higher income may have better healthy conditions because they have more opportunities to access good medications. We ignored those factors that can be regarded as confounding variables, which reduces the reliability of our results.
3. Causal inference cannot be drawn in our study. We conducted the t-test to test the sampling mean difference between two populations, and made conclusions that there are no causal connections between happiness & health and pet ownership based on the conditions of the occurrence of scores.

- Further questions, next steps

How do we minimize the sampling bias?
How do we minimize the response bias ?
How do we minimize the confounding bias?
How do we build the causal inference?
Do we have any recommendations in next steps?

1. One of the most effective methods that can be used to avoid sampling bias is simple random sampling, in which samples are chosen strictly by chance. We need to enlarge the sample sizes and share the survey via various methods such as email and website.
2. Response bias is a common form of information bias. We need to use more descriptive languages to define the happiness and health in our questionnaire, so people who take the survey can distinguish them neutrally. We have to make sure that all answer options are not leading, and check for alternative explanations if it is possible.
3. Confounding bias occurs when we do not account for all relevant effects. We can introduce control variables to control for confounding variables. For example, we could design questions about medical insurance, current income, and the happiest

experience to analyze people`s current mental state. However, successful adjustments for confounding require that large and random samples have been collected and are available for statistical analysis. Thus, confounding bias also needs be considered in the initial stage of a study.

4. When planning an observational study, we need to consider potential validity issues carefully, and plan observation and data collection in a way that minimizes sampling and response bias. After the well-designed study, we could build the causal inferences from an effective t-test that the p-value is less 0.05, and we reject the null hypothesis, building the causal connections between pet ownership and happiness &health.
5. A good designed observational study will have an important role in providing the information needed to improve the decision-making. There is always room for improvement and the hope that the future will bring better methods to further reduce the uncertainty and to increase the validity of its results.

Appendix

- The Knitted R markdown file
- The excel file

Stat Mid2 Project

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#data processing
#load the data; load the data in R
#Statistical Analysis
#data processing
#load the data; load the data in R
library(readxl)
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(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v readr   1.3.1
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

```
## -- Conflicts ---------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
# xlsx files

my_data2 <- read_excel("stat_data2.xlsx")
```

```
## New names:
## * `` -> ...5
## * `` -> ...7
## * `` -> ...21
```

```
head(my_data2)
```

```
## # A tibble: 6 x 23
##   PetOwner PIndex PHappyScore PHealthyScore ...5  Ptime ...7  PMonthlyMoney
##      <dbl>  <dbl>       <dbl>         <dbl> <lgl> <dbl> <lgl>         <dbl>
## 1        1      1           6             8 NA      1.3 NA              100
## 2        0      2           8             8 NA      1.5 NA              150
## 3        0      3           8             8 NA      1   NA              120
## 4        0      4           7             6 NA      0.5 NA               80
## 5        1      5          10             8 NA      2   NA              150
## 6        1      6           8             9 NA      2   NA              200
## # ... with 15 more variables: PGenderIndex <dbl>, PGender <chr>,
## #   PAge <dbl>, PetTypeIndex <dbl>, PetType <chr>, `No Pet` <lgl>,
## #   NIndex <dbl>, NHappyScore <dbl>, NHealthyScore <dbl>,
## #   NGenderIndex <dbl>, NGender <chr>, NAge <dbl>, ...21 <lgl>,
## #   Happy <dbl>, Healthy <dbl>
```

```
summary(my_data2)
```

```
##     PetOwner          PIndex       PHappyScore     PHealthyScore
##  Min.   :0.0000   Min.   : 1.00   Min.   : 5.000   Min.   : 4.000
##  1st Qu.:0.0000   1st Qu.:10.25   1st Qu.: 7.000   1st Qu.: 7.000
##  Median :1.0000   Median :19.50   Median : 8.000   Median : 7.000
##  Mean   :0.6909   Mean   :19.50   Mean   : 7.658   Mean   : 7.368
##  3rd Qu.:1.0000   3rd Qu.:28.75   3rd Qu.: 8.000   3rd Qu.: 8.000
##  Max.   :1.0000   Max.   :38.00   Max.   :10.000   Max.   :10.000
##                   NA's   :17      NA's   :17       NA's   :17
##    ...5             Ptime         ...7         PMonthlyMoney
##  Mode:logical   Min.   :0.100   Mode:logical   Min.   :  15.0
##  NA's:55        1st Qu.:0.500   NA's:55        1st Qu.: 100.0
##                 Median :1.000                  Median : 177.5
##                 Mean   :1.033                  Mean   : 194.9
##                 3rd Qu.:1.275                  3rd Qu.: 200.0
##                 Max.   :3.000                  Max.   :1000.0
##                 NA's   :17                     NA's   :17
##   PGenderIndex     PGender               PAge        PetTypeIndex
##  Min.   :1.000   Length:55          Min.   :23.00   Min.   :1.000
##  1st Qu.:1.000   Class :character   1st Qu.:25.00   1st Qu.:1.000
##  Median :2.000   Mode  :character   Median :28.00   Median :1.000
##  Mean   :1.579                      Mean   :31.39   Mean   :1.289
##  3rd Qu.:2.000                      3rd Qu.:34.00   3rd Qu.:2.000
##  Max.   :2.000                      Max.   :69.00   Max.   :2.000
##  NA's   :17                         NA's   :17      NA's   :17
##    PetType            No Pet            NIndex    NHappyScore
##  Length:55          Mode:logical   Min.   : 1   Min.   :6.000
##  Class :character   NA's:55        1st Qu.: 5   1st Qu.:7.000
##  Mode  :character                  Median : 9   Median :8.000
##                                    Mean   : 9   Mean   :7.647
##                                    3rd Qu.:13   3rd Qu.:8.000
##                                    Max.   :17   Max.   :9.000
##                                    NA's   :38   NA's   :38
##  NHealthyScore     NGenderIndex     NGender               NAge
##  Min.   : 5.000   Min.   :1.000   Length:55          Min.   :21.00
##  1st Qu.: 7.000   1st Qu.:1.000   Class :character   1st Qu.:24.00
##  Median : 8.000   Median :1.000   Mode  :character   Median :27.00
##  Mean   : 7.765   Mean   :1.471                      Mean   :27.35
##  3rd Qu.: 8.000   3rd Qu.:2.000                      3rd Qu.:29.00
##  Max.   :10.000   Max.   :2.000                      Max.   :40.00
##  NA's   :38       NA's   :38                         NA's   :38
##   ...21             Happy           Healthy
##  Mode:logical   Min.   : 5.000   Min.   : 4.000
##   NA's:55         1st Qu.: 7.000   1st Qu.: 7.000
##                   Median : 8.000   Median : 7.000
##                   Mean   : 7.655   Mean   : 7.491
##                   3rd Qu.: 8.000   3rd Qu.: 8.000
##                   Max.   :10.000   Max.   :10.000
##
```

```
#Exploratory Analysis
# sample size
n1 <- length(my_data2$PHappyScore[!is.na(my_data2$PHappyScore)])
n1
```

```
## [1] 38
```

```
n2 <- length(my_data2$PHealthyScore[!is.na(my_data2$PHappyScore)])
n2
```

```
## [1] 38
```

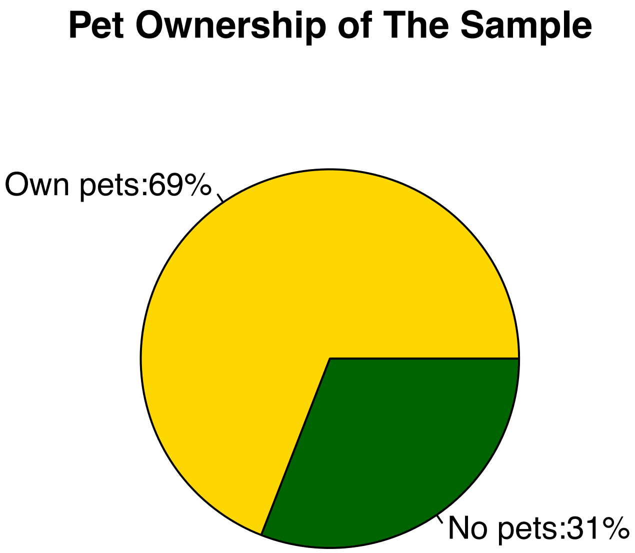
```
n3 <- length(my_data2$NHappyScore[!is.na(my_data2$NHappyScore)])
n3
```

```
## [1] 17
```

```
n4 <- length(my_data2$NHealthyScore[!is.na(my_data2$NHealthyScore)])
n4
```

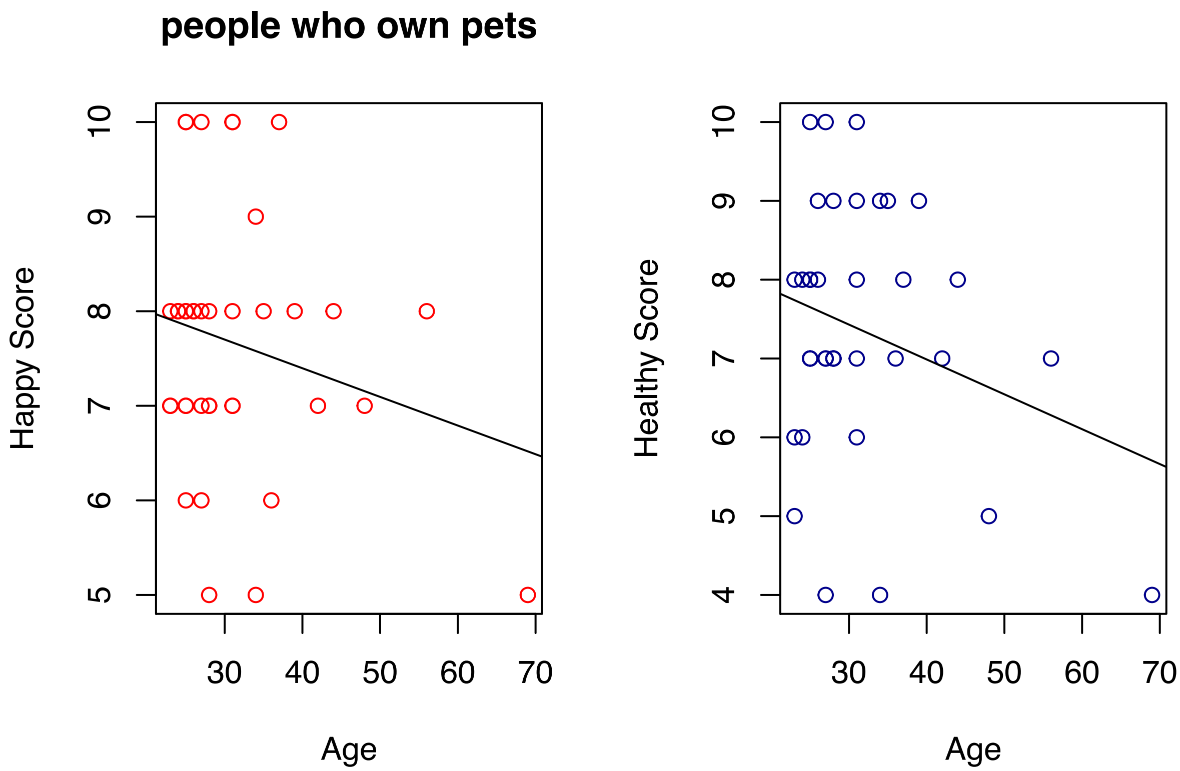
```
## [1] 17
```

```
# the pie chart
counts <- c(38,17)
lbls <- c("Own pets","No pets")
label <- paste(lbls,":",round(counts/sum(counts)*100), "%", sep="")
pie(counts,labels= label, col=c("gold","darkgreen"),
    main="Pet Ownership of The Sample")
```

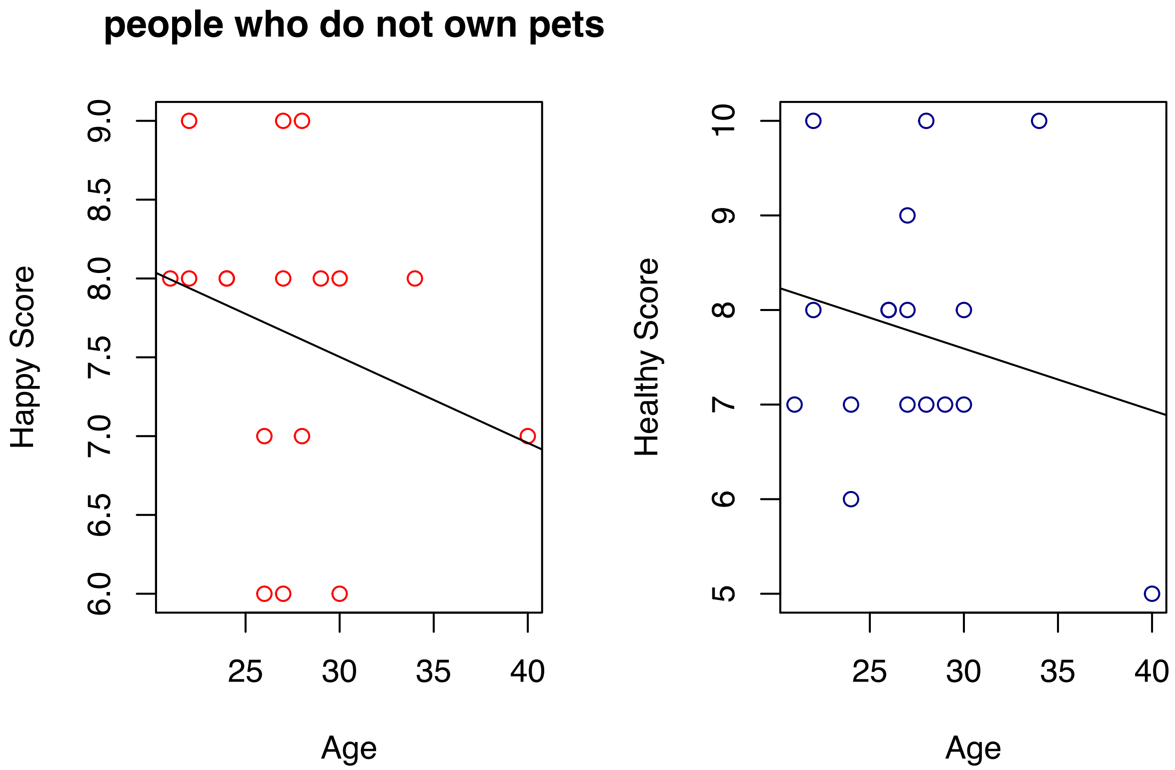


```
#scatter plot
par(mfrow=c(1,2))
plot (my_data2$PAge,my_data2$PHappyScore,main="people who own pets",
      xlab="Age", ylab="Happy Score", col='red')
abline(lm(my_data2$PHappyScore~my_data2$PAge ))
```

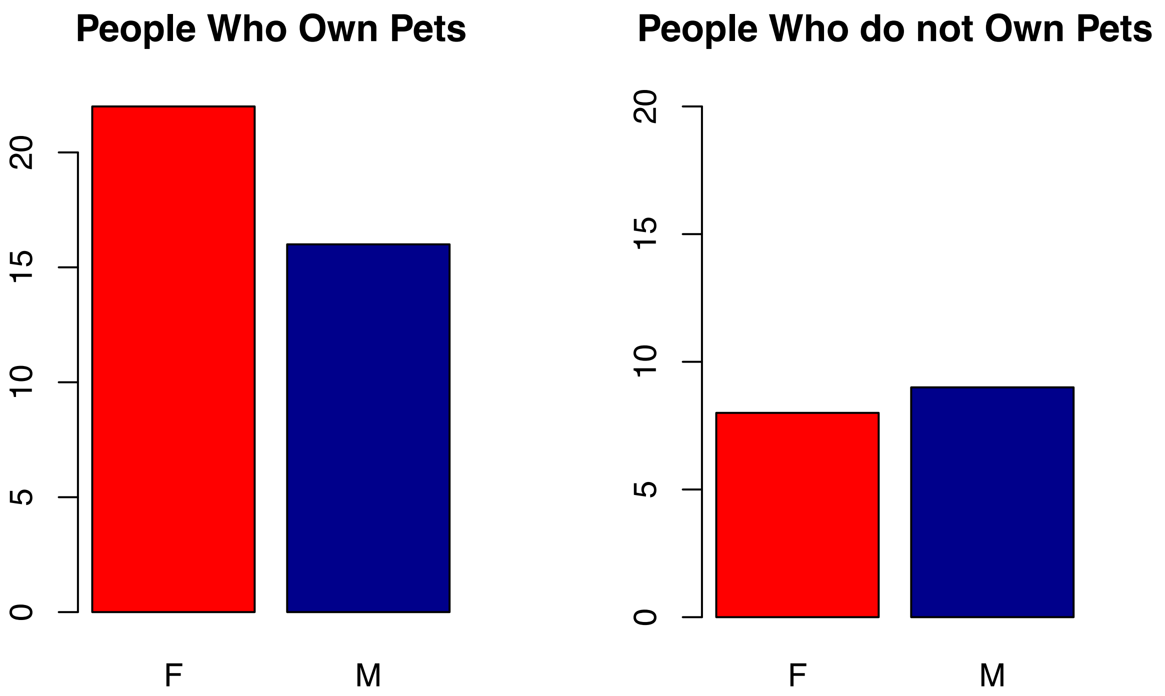
```
plot(my_data2$PAge,my_data2$PHealthyScore,
     xlab="Age", ylab="Healthy Score", col='darkblue')
abline(lm(my_data2$PHealthyScore~my_data2$PAge))
```



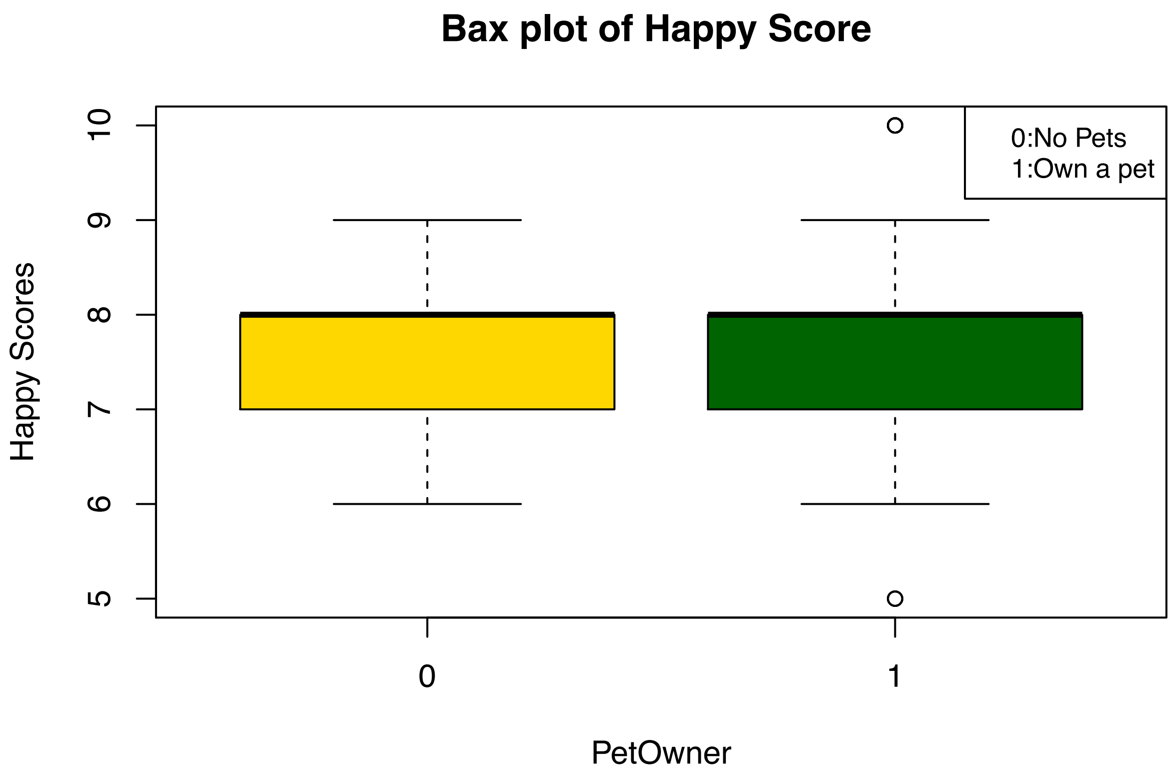
```
par(mfrow=c(1,2))
plot(my_data2$NAge,my_data2$NHappyScore,main=" people who do not own pets",
     xlab="Age", ylab="Happy Score", col='red')
abline(lm(my_data2$NHappyScore~my_data2$NAge))

plot(my_data2$NAge,my_data2$NHealthyScore,
     xlab="Age", ylab="Healthy Score", col='darkblue')
abline(lm(my_data2$NHealthyScore~my_data2$NAge))
```



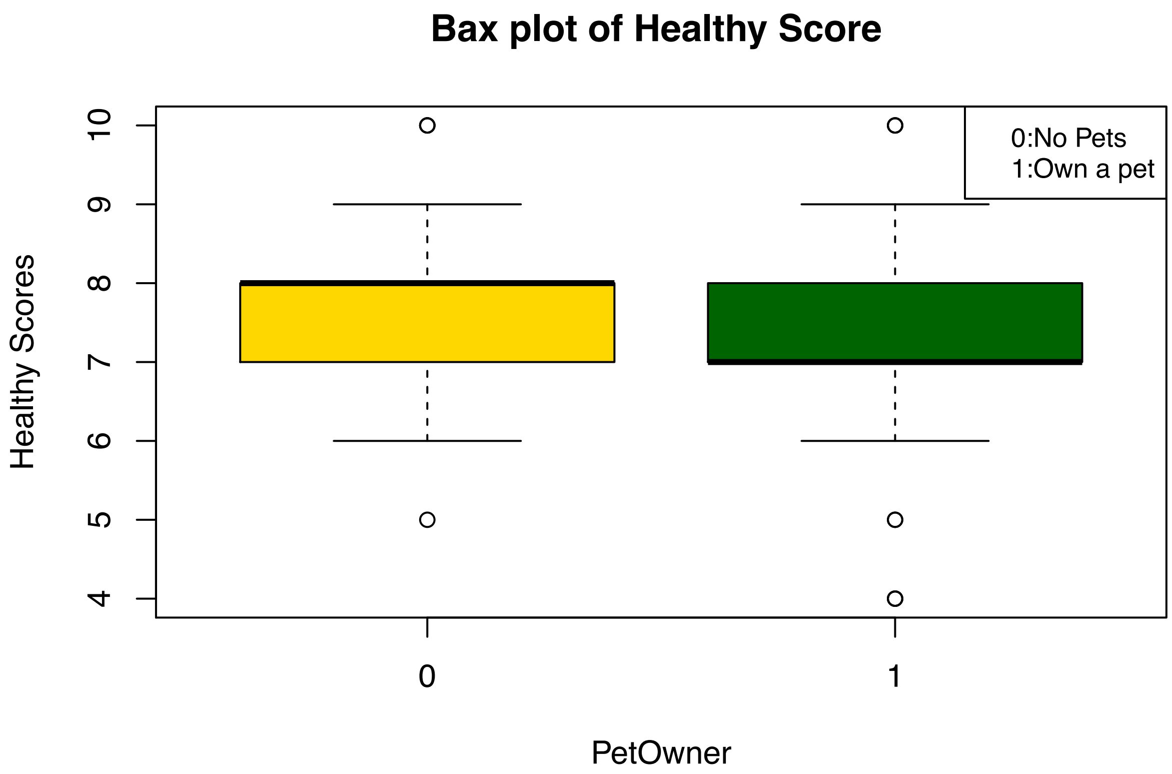
```
#barplot
par(mfrow=c(1,2))

counts1 <- table(my_data2$PGender)
barplot(counts1,
  main="People Who Own Pets",
  col=c("red","darkblue"))

counts2 <- table(my_data2$NGender)
barplot(counts2,
  main="People Who do not Own Pets",
  ylim=c(0,20),
  col=c("red","darkblue"))
```



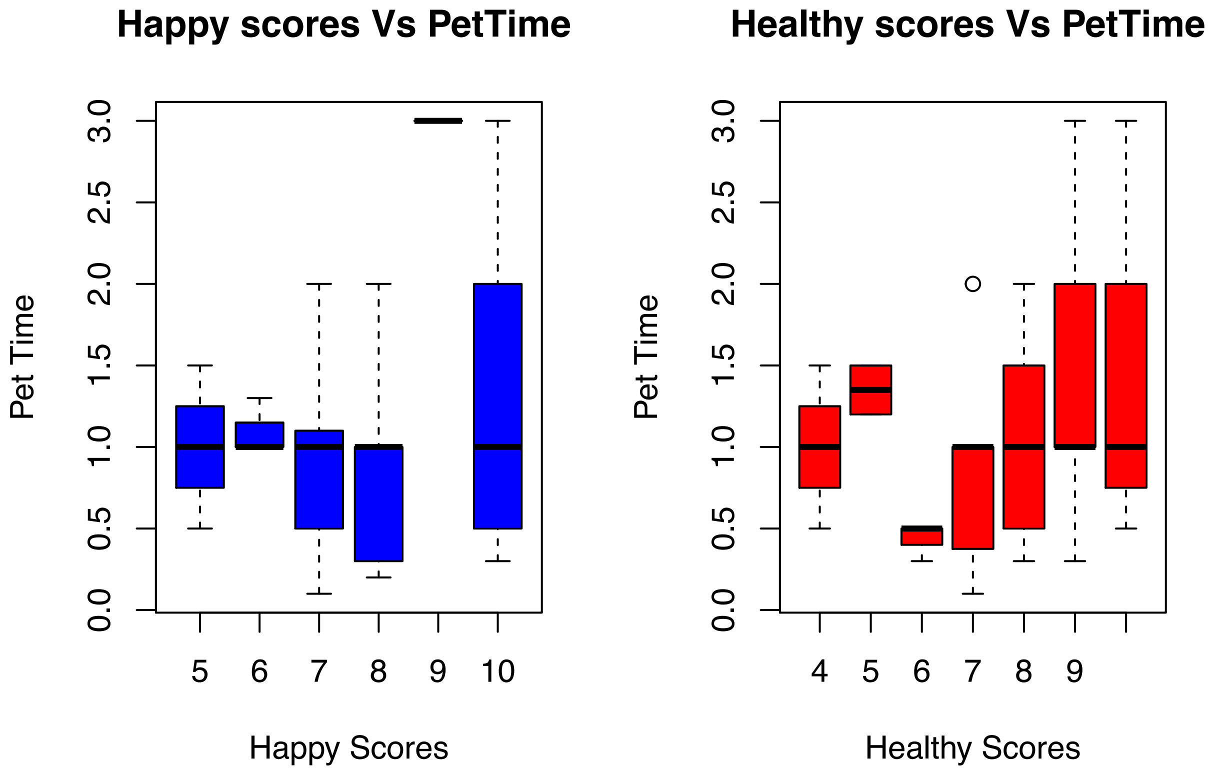
```
#Box-Plot
plot1 <- boxplot(Happy~PetOwner,
     data=my_data2,
     main="Bax plot of Happy Score ",
     ylab = "Happy Scores",
     col=c("gold","darkgreen"),
     border="black"
     )
legend("topright", legend = c("0:No Pets", "1:Own a pet"),cex=0.8)
```



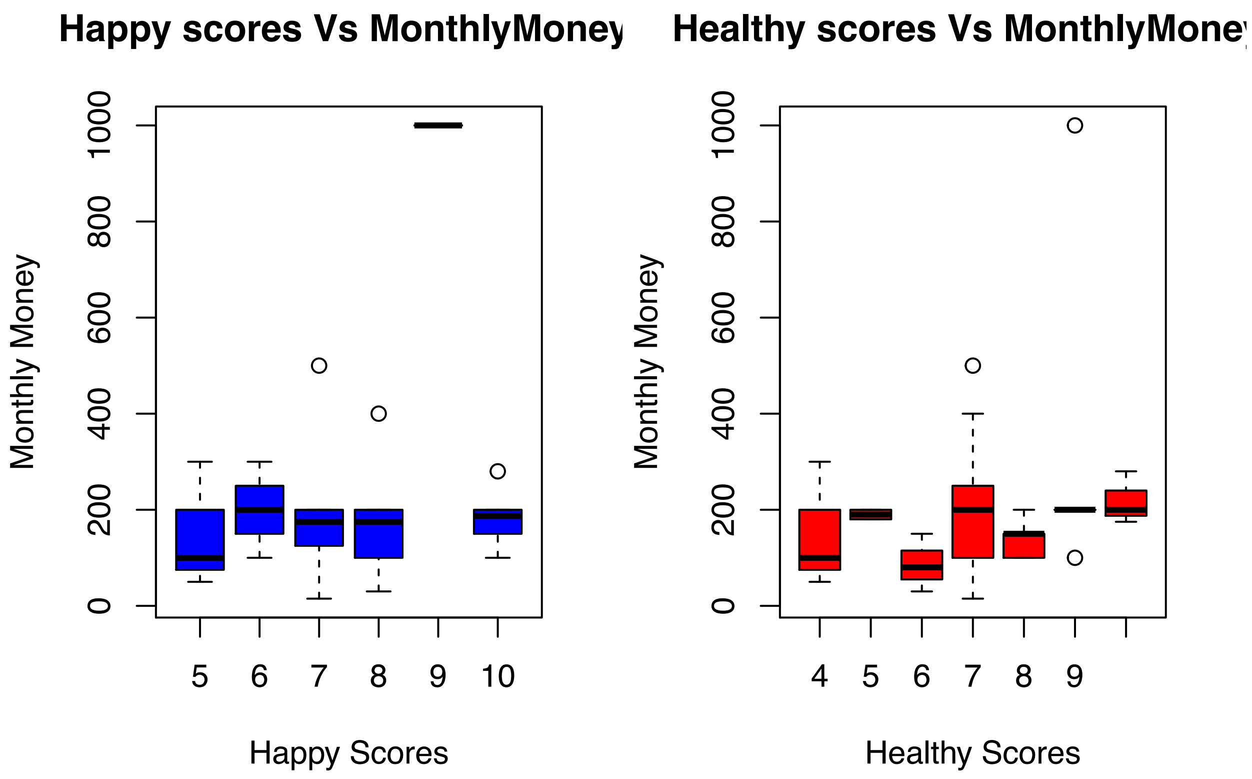
```
plot2 <- boxplot(Healthy~PetOwner,
     data=my_data2,
     main="Bax plot of Healthy Score ",
     ylab = "Healthy Scores",
     col=c("gold","darkgreen"),
     border="black"
     )
legend("topright", legend = c("0:No Pets", "1:Own a pet"),cex=0.8)
```



```r
par(mfrow=c(1,2))
# Happy scores Vs PetTime
plot1 <- boxplot(Ptime~PHappyScore,
     data=my_data2,
     main="Happy scores Vs PetTime ",
     xlab = "Happy Scores",
     ylab = "Pet Time",
     col="blue",
     border="black"
     )
# Healthy scores Vs PetTime
plot2 <- boxplot(Ptime~PHealthyScore,
     data=my_data2,
     main="Healthy scores Vs PetTime ",
     xlab = "Healthy Scores",
     ylab = "Pet Time",
     col="red",
     border="black"
     )
```



```
# Happy scores Vs MonthlyMoney
plot3 <- boxplot(PMonthlyMoney~PHappyScore,
     data=my_data2,
     main="Happy scores Vs MonthlyMoney ",
     xlab = "Happy Scores",
     ylab = "Monthly Money",
     col="blue",
     border="black"
     )
# Healthy scores Vs MonthlyMoney
plot4 <- boxplot(PMonthlyMoney~PHealthyScore,
     data=my_data2,
     main="Healthy scores Vs MonthlyMoney ",
     xlab = "Healthy Scores",
     ylab = "Monthly Money",
     col="red",
     border="black"
     )
```



```
#sample means
mean(my_data2$PHappyScore,na.rm=TRUE)
```

```
## [1] 7.657895
```

```
mean(my_data2$PHealthyScore,na.rm=TRUE)
```

```
## [1] 7.368421
```

```
mean(my_data2$NHappyScore, na.rm=TRUE)
```

```
## [1] 7.647059
```

```
mean(my_data2$NHealthyScore,na.rm=TRUE)
```

```
## [1] 7.764706
```

```
#sample st.dev
sd(my_data2$PHappyScore,na.rm=TRUE)
```

```
## [1] 1.380879
```

```
sd(my_data2$PHealthyScore,na.rm=TRUE)
```

```
## [1] 1.58406
```

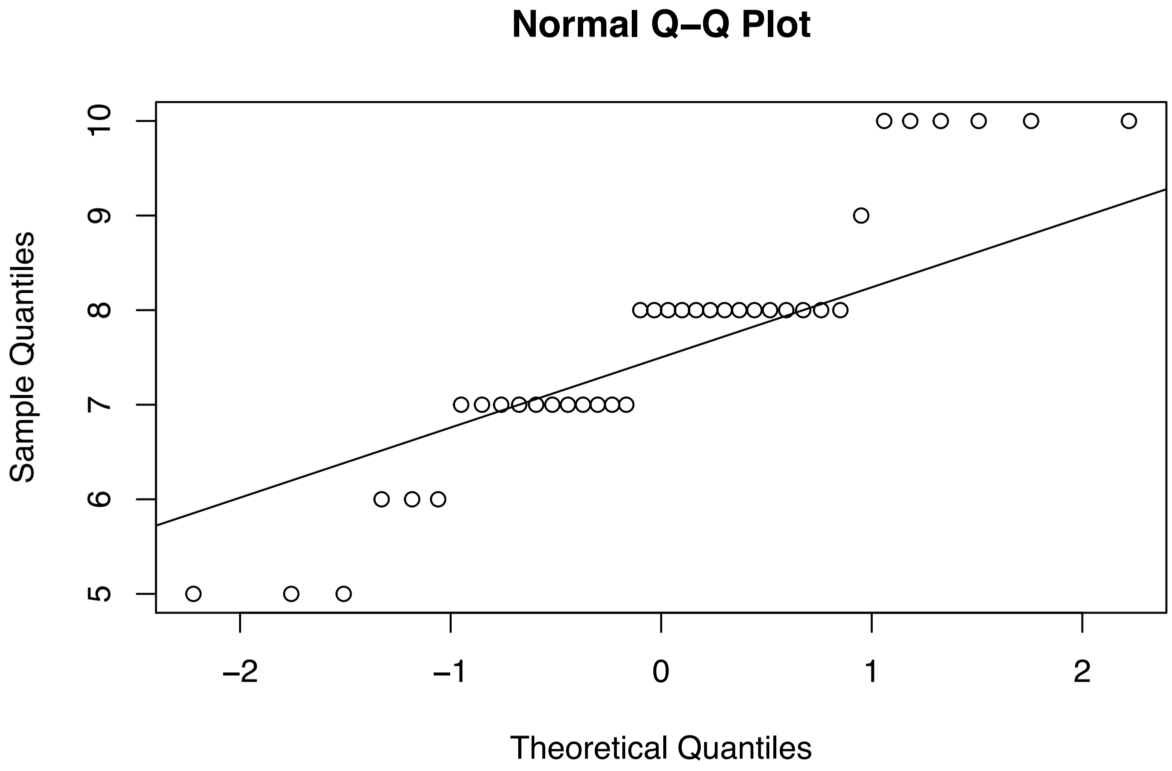
```
sd(my_data2$NHappyScore, na.rm=TRUE)
```

```
## [1] 0.9963167
```

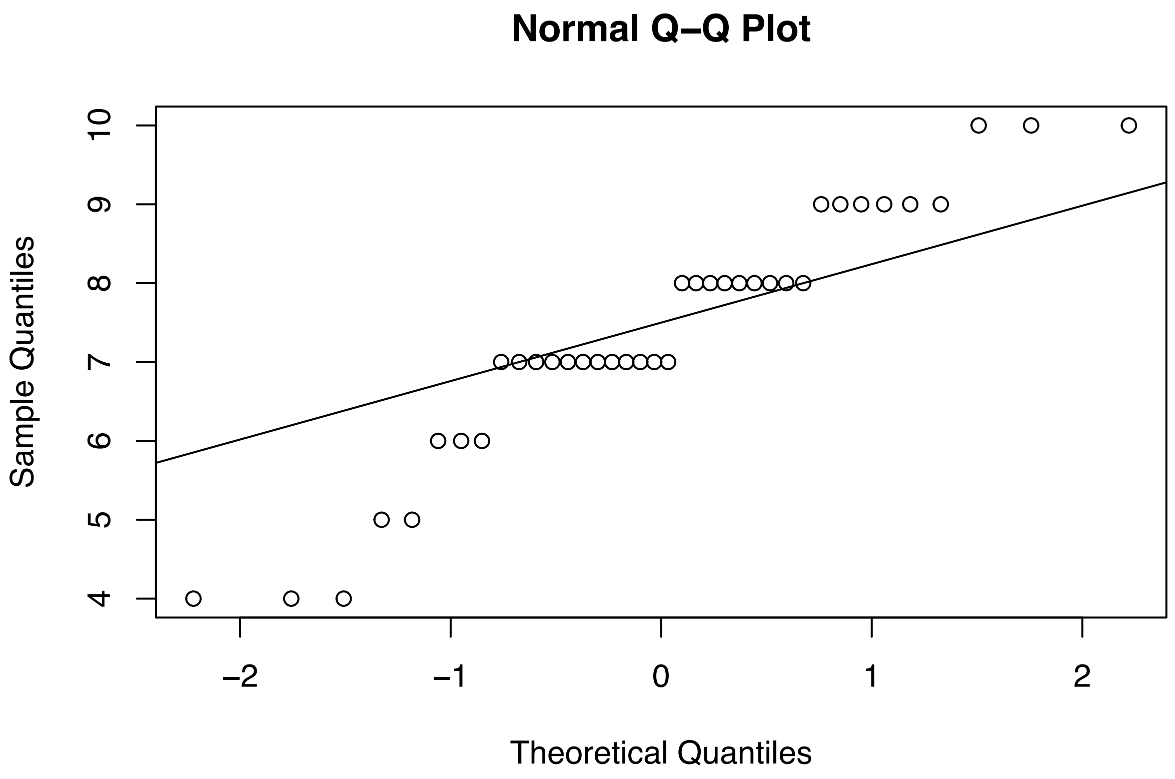
```
sd(my_data2$NHealthyScore,na.rm=TRUE)
```

```
## [1] 1.393261
```

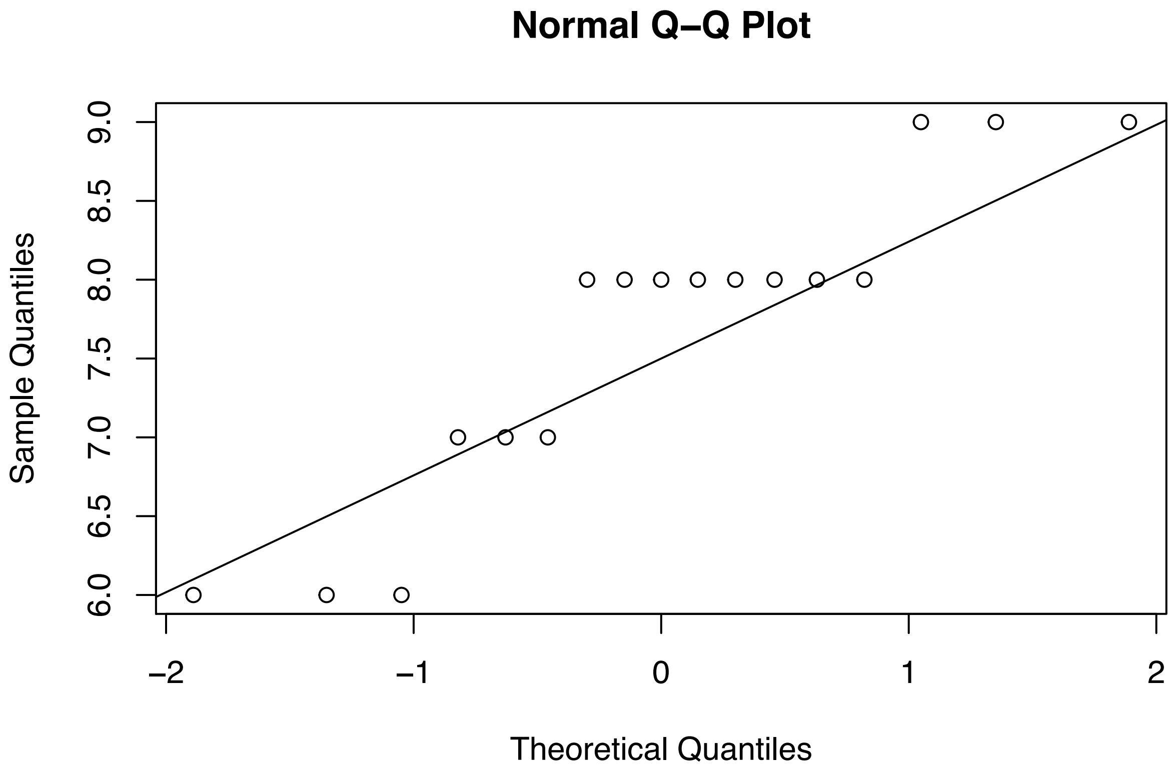
```
#check the population data is normally distributed
#check the qqplot firstly
qqnorm(my_data2$PHappyScore)
qqline(my_data2$PHappyScore)
```



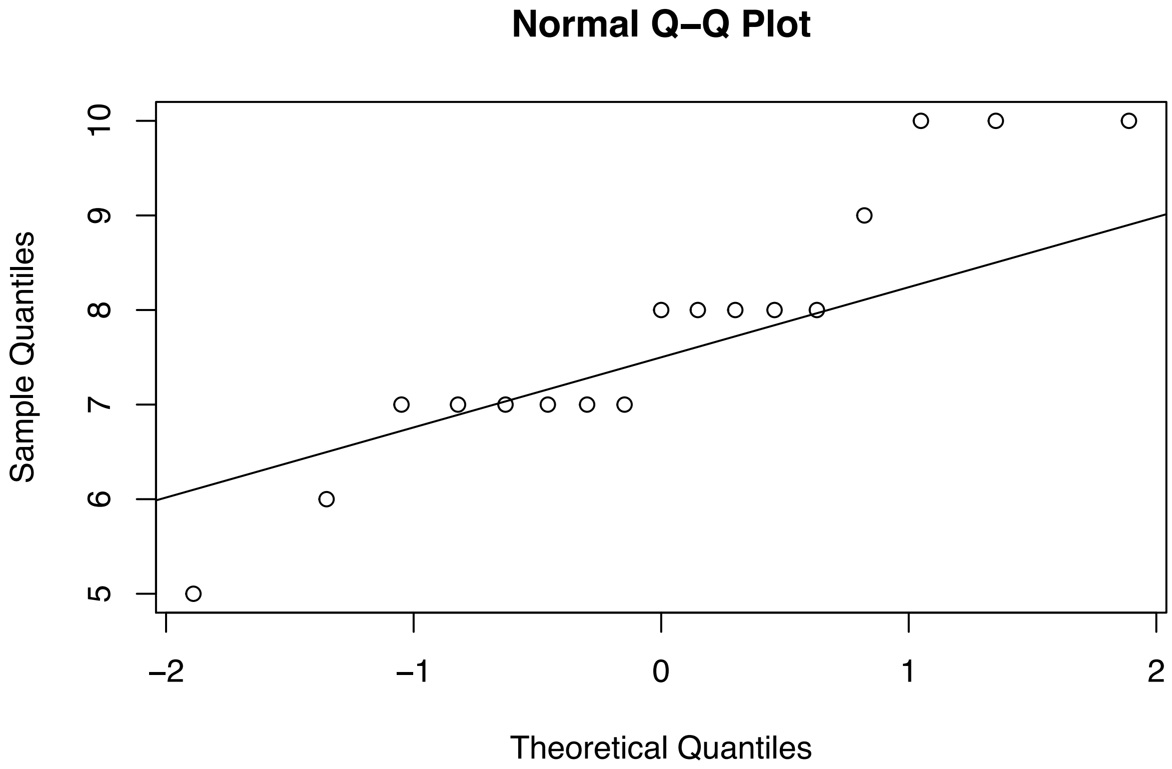
```
qqnorm(my_data2$PHealthyScore)
qqline(my_data2$PHealthyScore)
```



```
qqnorm(my_data2$NHappyScore)
qqline(my_data2$NHappyScore)
```



```
qqnorm(my_data2$NHealthyScore)
qqline(my_data2$NHealthyScore)
```



```
#order the data from the smallest to largest
#calculate the quantile
order1 <- order(my_data2$PHappyScore)
quantile(order1,seq(0.01,0.99,0.1))
```

```
##    1%   11%   21%   31%   41%   51%   61%   71%   81%   91%
##  1.54  6.94 12.34 17.74 23.14 28.54 33.94 39.34 44.74 50.14
```

```
order2 <- order(my_data2$PHealthyScore)
quantile(order2,seq(0.01,0.99,0.1))
```

```
##    1%   11%   21%   31%   41%   51%   61%   71%   81%   91%
##  1.54  6.94 12.34 17.74 23.14 28.54 33.94 39.34 44.74 50.14
```

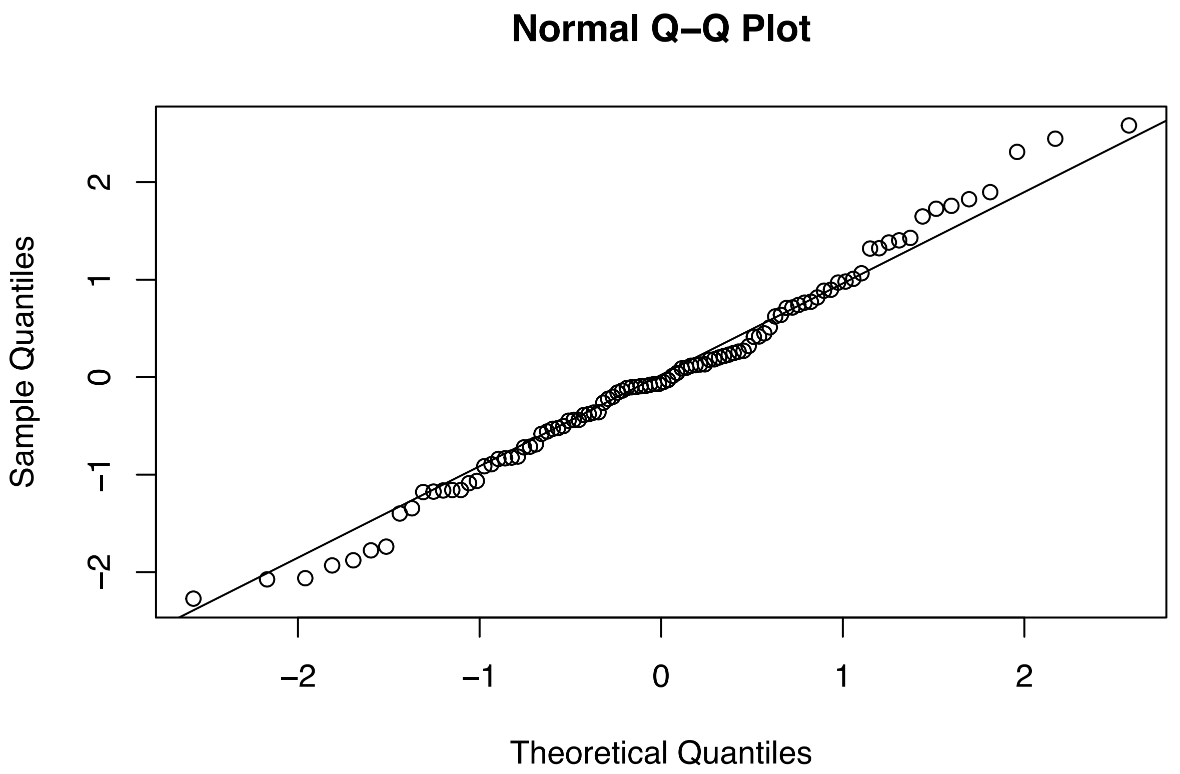
```
order3 <- order(my_data2$NHappyScore)
quantile(order2,seq(0.01,0.99,0.1))
```

```
##    1%   11%   21%   31%   41%   51%   61%   71%   81%   91%
##  1.54  6.94 12.34 17.74 23.14 28.54 33.94 39.34 44.74 50.14
```

```
order4 <- order(my_data2$NHealthyScore)
quantile(order2,seq(0.01,0.99,0.1))
```

```
##    1%   11%   21%   31%   41%   51%   61%   71%   81%   91%
##  1.54  6.94 12.34 17.74 23.14 28.54 33.94 39.34 44.74 50.14
```

```
#Normal distrbution
set.seed(100)
x <- rnorm(100)
qqnorm(x)
qqline(x)
```



```
quantile(x,seq(0.01,0.99,0.1))
```

```
##          1%         11%         21%         31%         41%         51% 
## -2.07637996 -1.16369702 -0.81666399 -0.44112017 -0.14611953 -0.03986303 
##         61%         71%         81%         91% 
##  0.18044155  0.42648148  0.83055240  1.40546230
```

```
#Two-sided t-test with 95% confidence interval (alpha=0.05)
#Happy
hypoth1 <- t.test(my_data2$PHappyScore,my_data2$NHappyScore,
                  data=my_data2,var.equal=TRUE)
hypoth1
```

```
## 
##  Two Sample t-test
## 
## data:  my_data2$PHappyScore and my_data2$NHappyScore
## t = 0.02908, df = 53, p-value = 0.9769
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7365551  0.7582269
## sample estimates:
## mean of x mean of y 
##  7.657895  7.647059
```

```
hypoth1$conf.int
```

```
## [1] -0.7365551  0.7582269
## attr(,"conf.level")
## [1] 0.95
```

```
hypoth1$p.value
```

```
## [1] 0.9769101
```

```
#Healthy
hypoth2 <- t.test(my_data2$PHealthyScore,my_data2$NHealthyScore,
                  data=my_data2,var.equal=TRUE)
hypoth2
```

```
##
##  Two Sample t-test
##
## data:  my_data2$PHealthyScore and my_data2$NHealthyScore
## t = -0.88827, df = 53, p-value = 0.3784
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2911145  0.4985448
## sample estimates:
## mean of x mean of y
##  7.368421  7.764706
```

```
hypoth2$conf.int
```

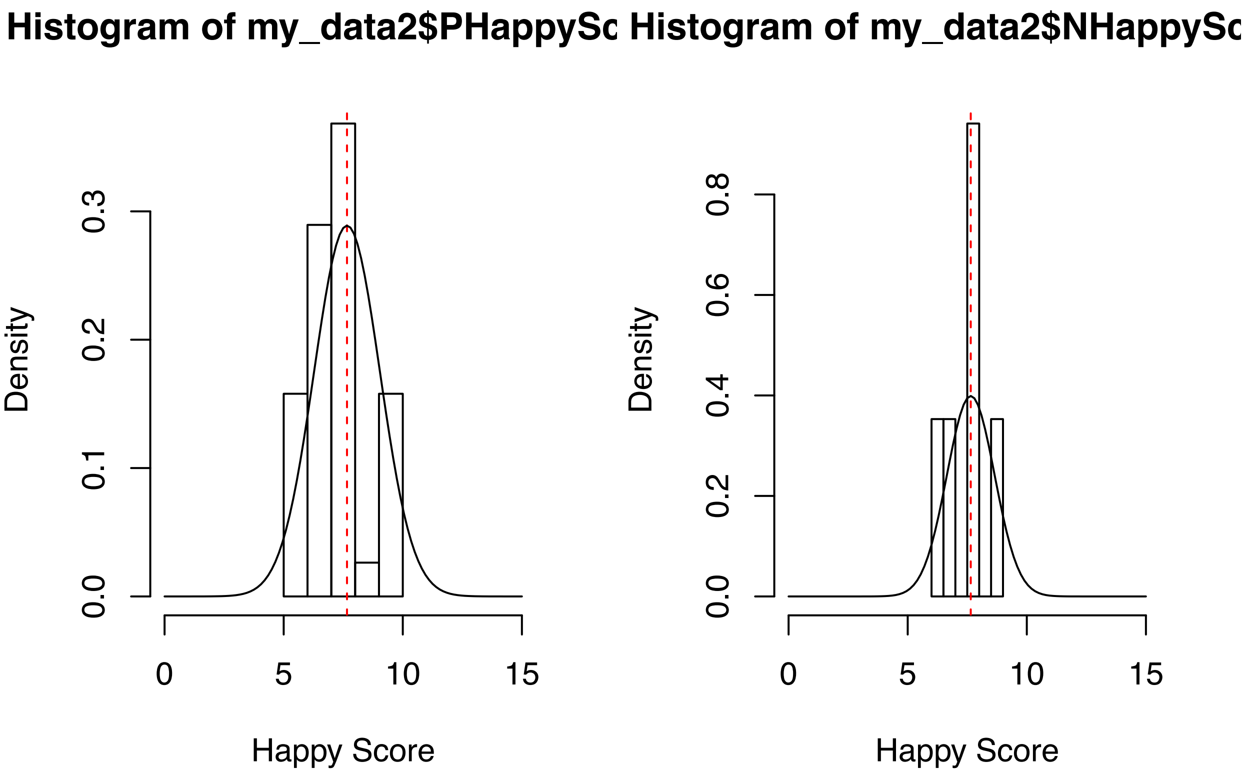
```
## [1] -1.2911145  0.4985448
## attr(,"conf.level")
## [1] 0.95
```

```
hypoth2$p.value
```

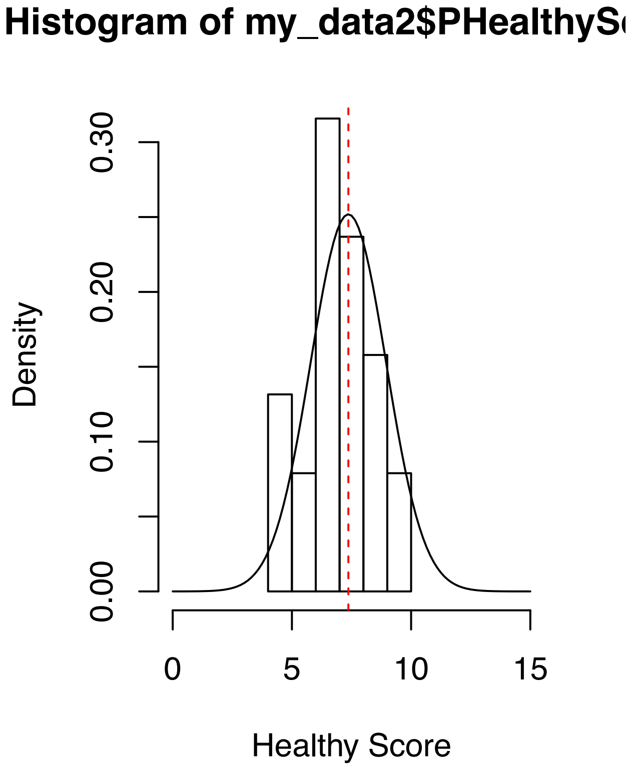
```
## [1] 0.3784114
```

```
#Plot a histogram of the sampling distribution with normal distribution curve
#people have pets
par(mfrow=c(1,2))

h1 <-hist(my_data2$PHappyScore,
    xlab="Happy Score",xlim=c(0,15),ylab="Density", freq=FALSE)
abline(v=mean(my_data2$PHappyScore,na.rm=TRUE),col="red",lty="dashed")
curve(dnorm(x, mean=mean(my_data2$PHappyScore,na.rm=TRUE),
            sd=sd(my_data2$PHappyScore,na.rm=TRUE)), add=TRUE, col="black")


h2 <-hist(my_data2$NHappyScore,
     xlab="Happy Score",xlim=c(0,15),ylab="Density", freq=FALSE)
abline(v=mean(my_data2$NHappyScore,na.rm=TRUE),col="red",lty="dashed")
curve(dnorm(x, mean=mean(my_data2$NHappyScore, na.rm=TRUE,)),
      sd=sd(mean(my_data2$NHappyScore, na.rm=TRUE)), add=TRUE, col="black")
```

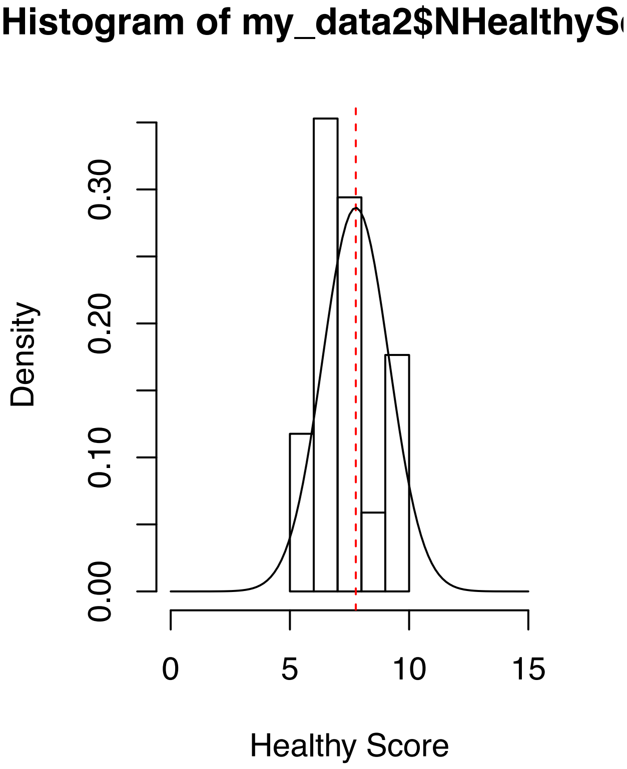
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "sd" is not a
## graphical parameter
```



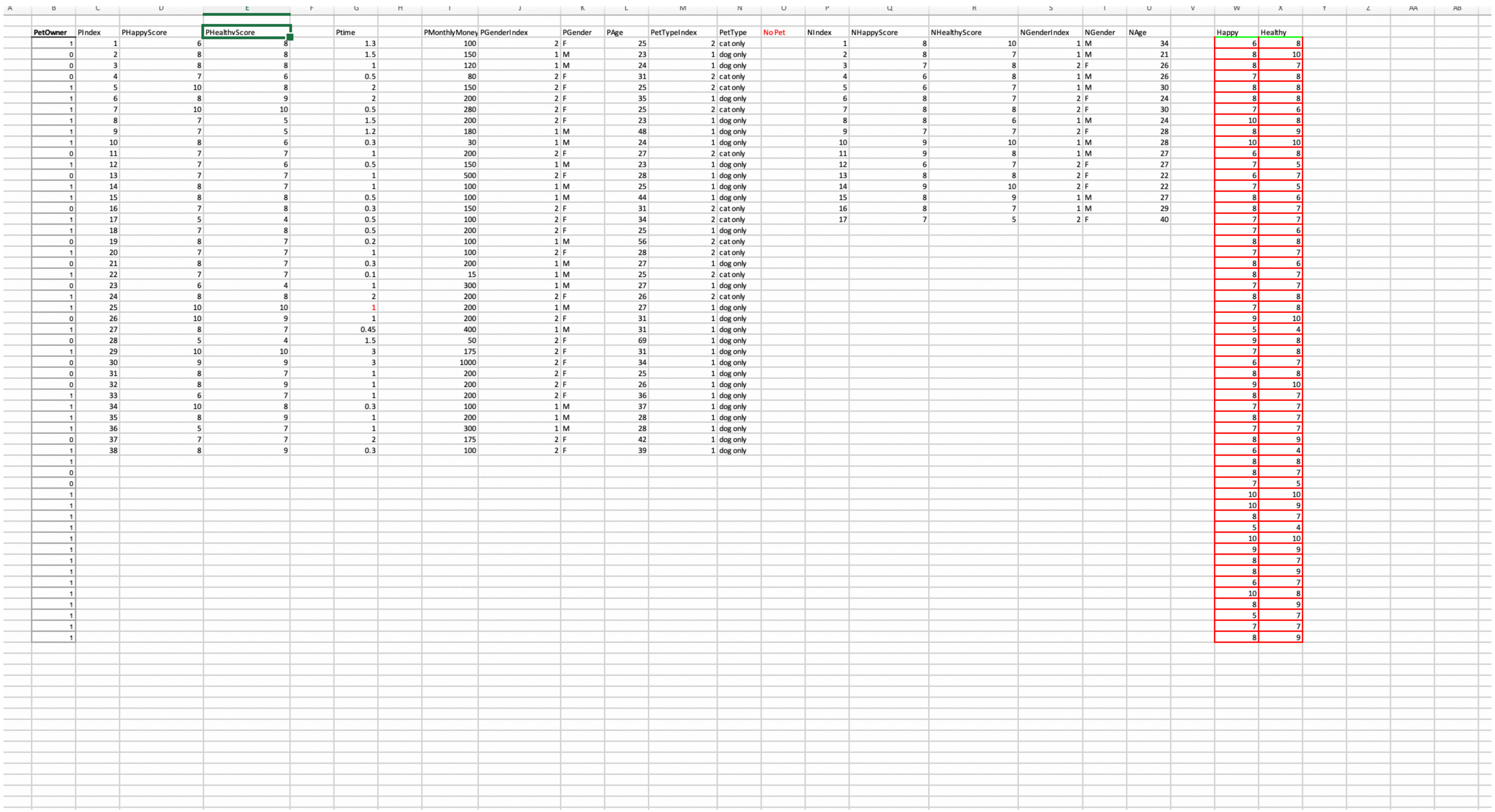
```
#people do not have pets
par(mfrow=c(1,2))

h3 <- hist(my_data2$PHealthyScore,xlab="Healthy Score",xlim=c(0,15),
           ylab="Density", freq=FALSE)
abline(v=mean(my_data2$PHealthyScore,na.rm=TRUE),col="red",lty="dashed")
curve(dnorm(x, mean=mean(my_data2$PHealthyScore,na.rm=TRUE),
            sd=sd(my_data2$PHealthyScore,na.rm=TRUE)), add=TRUE, col="black")

hist(my_data2$NHealthyScore,
     xlab="Healthy Score",xlim=c(0,15),ylab="Density", freq=FALSE)
abline(v=mean(my_data2$NHealthyScore,na.rm=TRUE),col="red",lty="dashed")
curve(dnorm(x, mean=mean(my_data2$NHealthyScore,na.rm=TRUE),
            sd=sd(my_data2$NHealthyScore,na.rm=TRUE)), add=TRUE, col="black")
```



Histogram of my_data2$PHealthyS
Density
0.00
0.10
0.20
0.30
0
5
10
15
Healthy Score



Histogram of my_data2$NHealthyS
Density
0.00
0.10
0.20
0.30
0
5
10
15
Healthy Score



PetOwner	PIndex	PHappyScore	PHealthvScore	Ptime	PMonthlyMoney	PGenderIndex	PGender	PAge	PetTypeIndex	PetType	No Pet	NIndex	NHappyScore	NHealthyScore	NGenderIndex	NGender	NAge	Happy	Healthy
1	1	6	8	1.3	100	2	F	25	2	cat only		1	8	10	1	M	34	6	8
0	2	8	8	1.5	150	1	M	23	1	dog only		2	8	7	1	M	21	8	10
0	3	8	8	1	120	1	M	24	1	dog only		3	7	8	2	F	26	8	7
0	4	7	6	0.5	80	2	F	31	2	cat only		4	6	8	1	M	26	7	8
1	5	10	8	2	150	2	F	25	2	cat only		5	6	7	1	M	30	8	8
1	6	8	9	2	200	2	F	35	1	dog only		6	8	7	2	F	24	8	8
1	7	10	10	0.5	280	2	F	25	2	cat only		7	8	8	2	F	30	7	6
1	8	7	5	1.5	200	2	F	23	1	dog only		8	8	6	1	M	24	10	8
1	9	7	5	1.2	180	1	M	48	1	dog only		9	7	7	2	F	28	8	9
1	10	8	6	0.3	30	1	M	24	1	dog only		10	9	10	1	M	28	10	10
0	11	7	7	1	200	2	F	27	2	cat only		11	9	8	1	M	27	6	8
1	12	7	6	0.5	150	1	M	23	1	dog only		12	6	7	2	F	27	7	5
0	13	7	7	1	500	2	F	28	1	dog only		13	8	8	2	F	22	6	7
1	14	8	7	1	100	1	M	25	1	dog only		14	9	10	2	F	22	7	5
1	15	8	8	0.5	100	1	M	44	1	dog only		15	8	9	1	M	27	8	6
0	16	7	8	0.3	150	2	F	31	2	cat only		16	8	7	1	M	29	8	7
1	17	5	4	0.5	100	2	F	34	2	cat only		17	7	5	2	F	40	7	7
1	18	7	8	0.5	200	2	F	25	1	dog only								7	6
0	19	8	7	0.2	100	1	M	56	2	cat only								8	8
1	20	7	7	1	100	2	F	28	2	cat only								7	7
0	21	8	7	0.3	200	1	M	27	1	dog only								8	6
1	22	7	7	0.1	15	1	M	25	2	cat only								8	7
0	23	6	4	1	300	1	M	27	1	dog only								7	7
1	24	8	8	2	200	2	F	26	2	cat only								8	8
1	25	10	10	1	200	1	M	27	1	dog only								7	8
0	26	10	9	1	200	2	F	31	1	dog only								9	10
1	27	8	7	0.45	400	1	M	31	1	dog only								5	4
0	28	5	4	1.5	50	2	F	69	1	dog only								9	8
1	29	10	10	3	175	2	F	31	1	dog only								7	8
0	30	9	9	3	1000	2	F	34	1	dog only								6	7
0	31	8	7	1	200	2	F	25	1	dog only								8	8
0	32	8	9	1	200	2	F	26	1	dog only								9	10
1	33	6	7	1	200	2	F	36	1	dog only								8	7
1	34	10	8	0.3	100	1	M	37	1	dog only								7	7
1	35	8	9	1	200	1	M	28	1	dog only								8	7
1	36	5	7	1	300	1	M	28	1	dog only								7	7
0	37	7	7	2	175	2	F	42	1	dog only								8	9
1	38	8	9	0.3	100	2	F	39	1	dog only								6	4
1																		8	8
0																		8	7
0																		7	5
1																		10	10
1																		10	9
1																		8	7
1																		5	4
1																		10	10
1																		9	9
1																		8	7
1																		8	9
1																		6	7
1																		10	8
1																		8	9
1																		5	7
1																		7	7
1																		8	9