EDA_Toy_Dataset

The toys dataset is a fictional dataset avaliable on kaggle primiarly used to exploratory data analysis

The dataset is avaliabel at: https://www.kaggle.com/carlolepelaars/toy-dataset.

This Rmarkdown describes my EDA of the toys dataset.

Package Loading

```r
library(dplyr)
library(ggplot2)
library(ggpubr)
```

Dataset Loading

```r
toys <- read.csv("toy_dataset.csv")

toys <- toys %>% mutate(across(c("Illness","Gender","City"),as.factor))

summary(toys)
```

```
##      Number                  City          Gender           Age
##  Min.   :     1   New York City:50307   Female:66200   Min.   :25.00
##  1st Qu.: 37501   Los Angeles  :32173   Male  :83800   1st Qu.:35.00
##  Median : 75000   Dallas       :19707                  Median :45.00
##  Mean   : 75000   Mountain View:14219                  Mean   :44.95
##  3rd Qu.:112500   Austin       :12292                  3rd Qu.:55.00
##  Max.   :150000   Boston       : 8301                  Max.   :65.00
##                   (Other)      :13001
##      Income       Illness
##  Min.   :  -654   No :137861
##  1st Qu.: 80868   Yes: 12139
##  Median : 93655
##  Mean   : 91253
##  3rd Qu.:104519
##  Max.   :177157
##
```

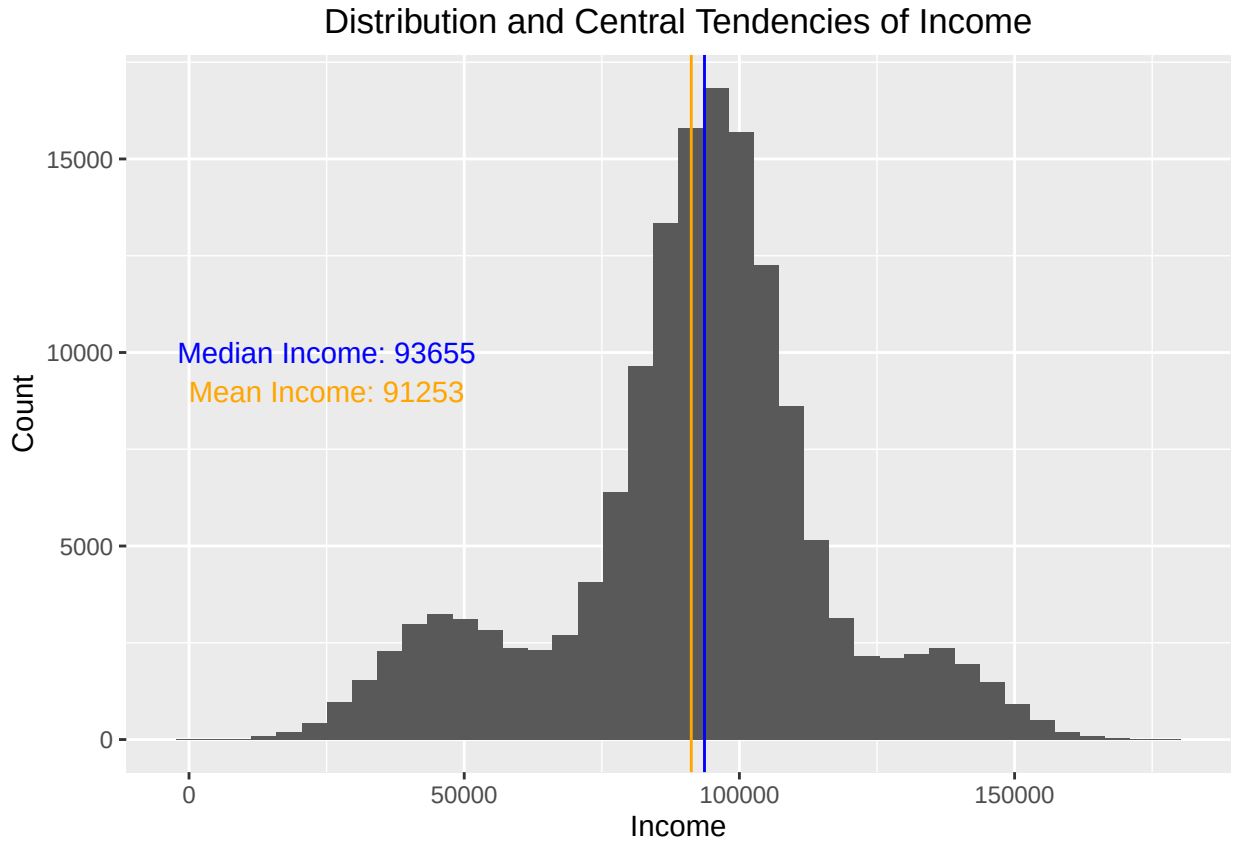
```r
out <- paste0("The toys dataset has ",nrow(toys), " rows and ",ncol(toys)," columns.")

out
```

```
## [1] "The toys dataset has 150000 rows and 6 columns."
```

The toy_dataset has no missing values but a column at the beginning that deontes that the observation number.

Initial Visualization of the Dataset

```
Income <- ggplot(toys,aes(x = Income)) +
  geom_histogram(bins = 40) +
  ylab("Count") +
  geom_vline(xintercept = round(median(toys$Income),0),color = "blue") +
  annotate(geom = "text",label = paste0("Mean Income: ",round(mean(toys$Income),0)),
           x = 25000,y=9000,color = "orange") +
  geom_vline(xintercept = round(mean(toys$Income),0),color = "orange") +
  annotate(geom = "text",label = paste0("Median Income: ",round(median(toys$Income),1)),
           x = 25000,y=10000,color ="blue") +
  labs(title = "Distribution and Central Tendencies of Income") +
  theme(plot.title = element_text(hjust = 0.5))

Income
```



Income grouped by City and Gender

```
Income_by_city_and_gender <-
ggplot(toys,aes(x = City,y = Income)) +
geom_boxplot(aes(color = Gender)) +
labs(title = "Median Income by City and Gender") +
theme(axis.text.x = element_text(vjust = 1,angle = 45,hjust = 1),
```

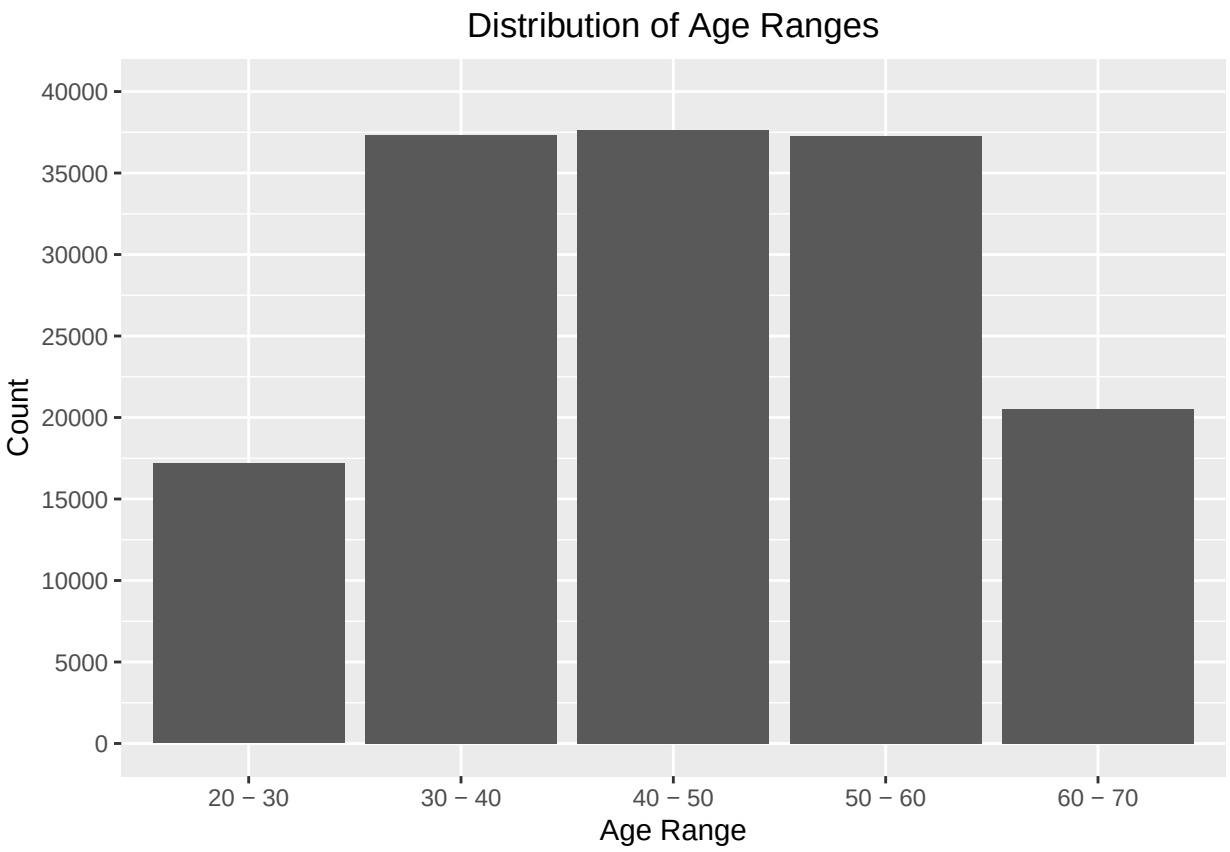
```
      plot.title = element_text(hjust = 0.5))

Income_by_city_and_gender
```

From the graph above, it appears that the median income is higher in men in women. This is consistent in all cities. The highest income amongst men and women is in Mountain View followed by major cities in the United States including Los Angele and New York City.

Visualization of Age distribution

```
Age <-
  toys %>%
  mutate(age = case_when(toys$Age >= 20 &  toys$Age < 30  ~ "20 - 30",
                         toys$Age >= 30 &  toys$Age < 40  ~ "30 - 40",
                         toys$Age >= 40 &  toys$Age < 50  ~ "40 - 50",
                         toys$Age >= 50 &  toys$Age < 60  ~ "50 - 60",
                         toys$Age >= 60 &  toys$Age < 70  ~ "60 - 70")) %>%
  group_by(age) %>% summarize(count = n()) %>%
  ggplot(.,aes(x = age,y = count)) +  geom_col() +
  scale_y_continuous(limits = c(0,40000),breaks = seq(0,40000,5000)) +
  labs(y = "Count",x = "Age Range",title= "Distribution of Age Ranges") +
  theme(plot.title = element_text(hjust = 0.5))
```

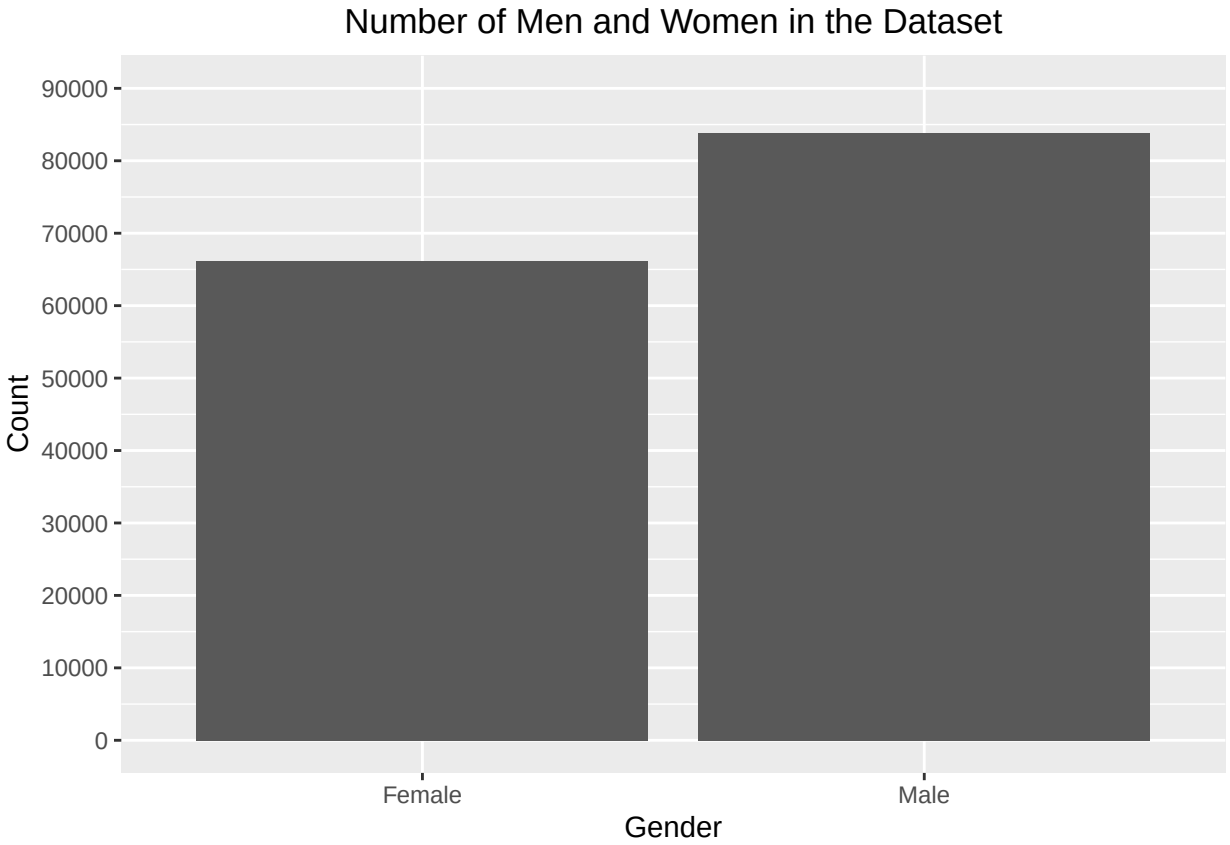
```
Age
```



The majority of people in the dataset are between the ages of 30 and 60. This is line with the mean age in the dataset being 45 years old.

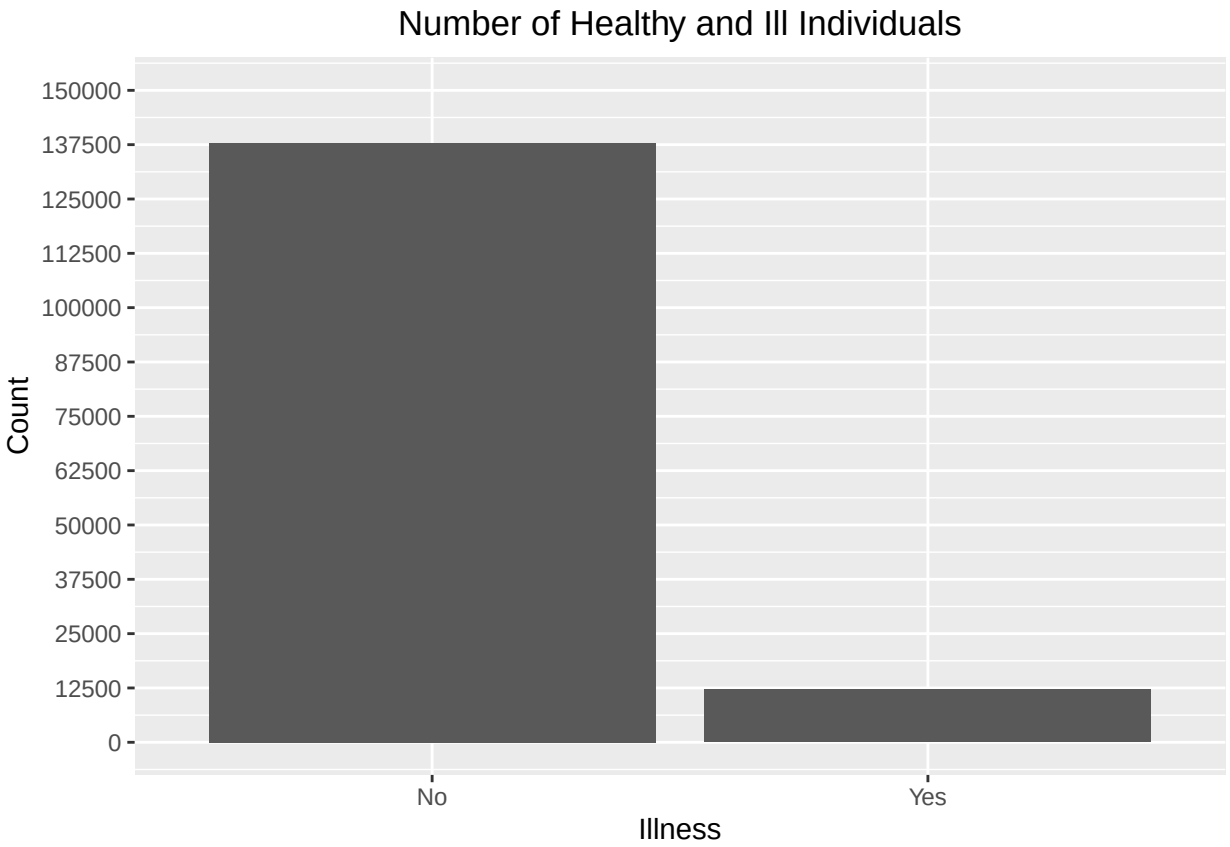
Total count of Gender and Illness in the dataset

```
Gender <-
  toys %>% group_by(Gender) %>% summarise(count = n()) %>%
  ggplot(.,aes(x = Gender, y = count)) + geom_bar(stat = "identity") +
  ylab("Count") + scale_y_continuous(limits = c(0,90000),
                                     breaks = seq(0,90000,10000)) +
  labs(title = "Number of Men and Women in the Dataset") +
  theme(plot.title = element_text(hjust = 0.5))

Gender
```



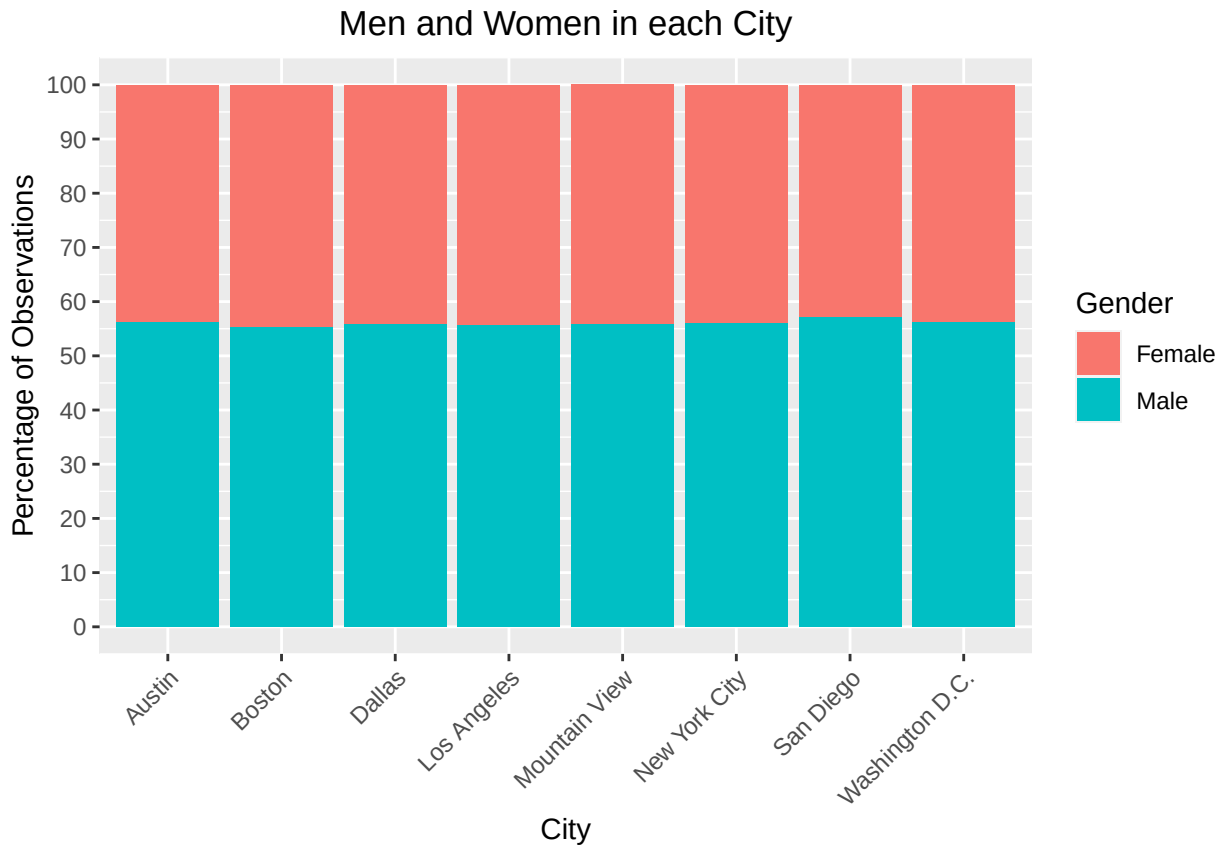
44.1% and 55.8% of this dataset is composed of men and women, respectively.

```
Illness <-
    toys %>% group_by(Illness) %>% summarise(count = n()) %>%
    ggplot(.,aes(x = Illness, y = count)) + geom_bar(stat = "identity") +
    scale_y_continuous(limits = c(0,150000),breaks = seq(0,150000,12500)) + ylab("Count") +
  labs(title = "Number of Healthy and Ill Individuals") +
  theme(plot.title = element_text(hjust = 0.5))

Illness
```



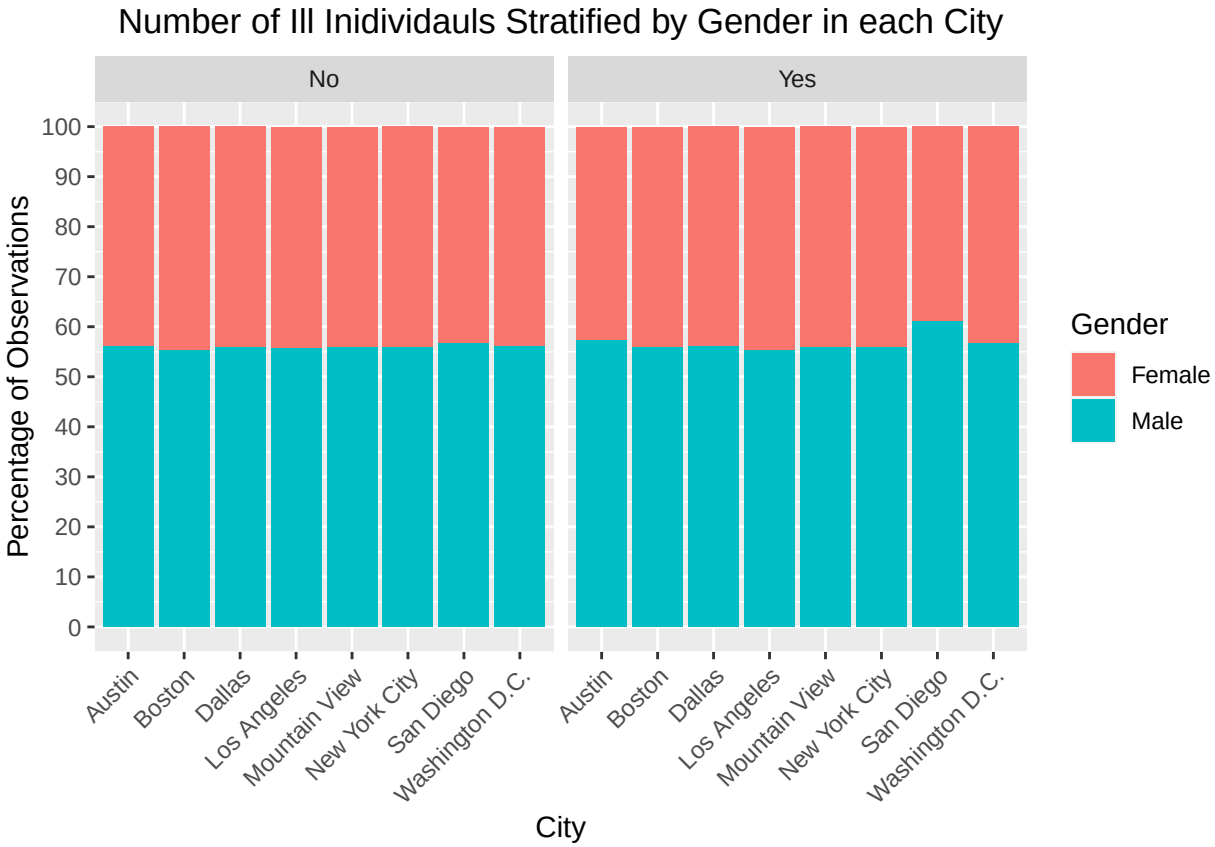
The toys dataset details the number of ill and healthy individuals. The majority of individuals in the dataset are healthy but 8.3% of the individuals are ill. This differs from the number of individuals who are male and female where both categories compose similar percentages of the observations in the dataset.

```
toys %>% group_by(City,Gender) %>%
  summarize(Count = n(),.groups = "drop") %>% group_by(City) %>%
  mutate(percentage = Count/sum(Count)) %>%
  ggplot(.,aes(x = City, y = percentage*100,fill = Gender)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(limits = c(0,100),breaks = seq(0,100,10)) +
  theme(axis.text.x = element_text(vjust = 1,angle = 45,hjust = 1),
        plot.title = element_text(hjust = 0.5)) +
  labs(title = "Men and Women in each City",y= "Percentage of Observations")
```



The percentage of men and women in the dataset is consitent across all cities at 55% and 45%, respectively. The male to female ratio in the cities is similar to that in the entire dataset.

```
toys %>% group_by(City,Illness,Gender) %>%
  summarize(Count = n(),.groups = "drop") %>% group_by(City,Illness) %>%
  mutate(percentage = Count/sum(Count)) %>%
  ggplot(.,aes(x = City, y = percentage*100,fill = Gender)) +
  geom_bar(stat = "identity") + facet_wrap(vars(Illness)) +
  scale_y_continuous(limits = c(0,100),breaks = seq(0,100,10)) +
  labs(title = "Number of Ill Inidividauls Stratified by Gender in each City",
       y  = "Percentage of Observations") +
  theme(axis.text.x = element_text(vjust = 1,angle = 45,hjust = 1),
        plot.title = element_text(hjust = 0.5))
```


Number of Ill Inidividauls Stratified by Gender in each City
No
Yes
Percentage of Observations
100
90
80
70
60
50
40
30
20
10
0
Austin
Boston
Dallas
Los Angeles
Mountain View
New York City
San Diego
Washington D.C.
City
Gender
Female
Male