Hmk4-Q3

The variable chosen is ‘income’.

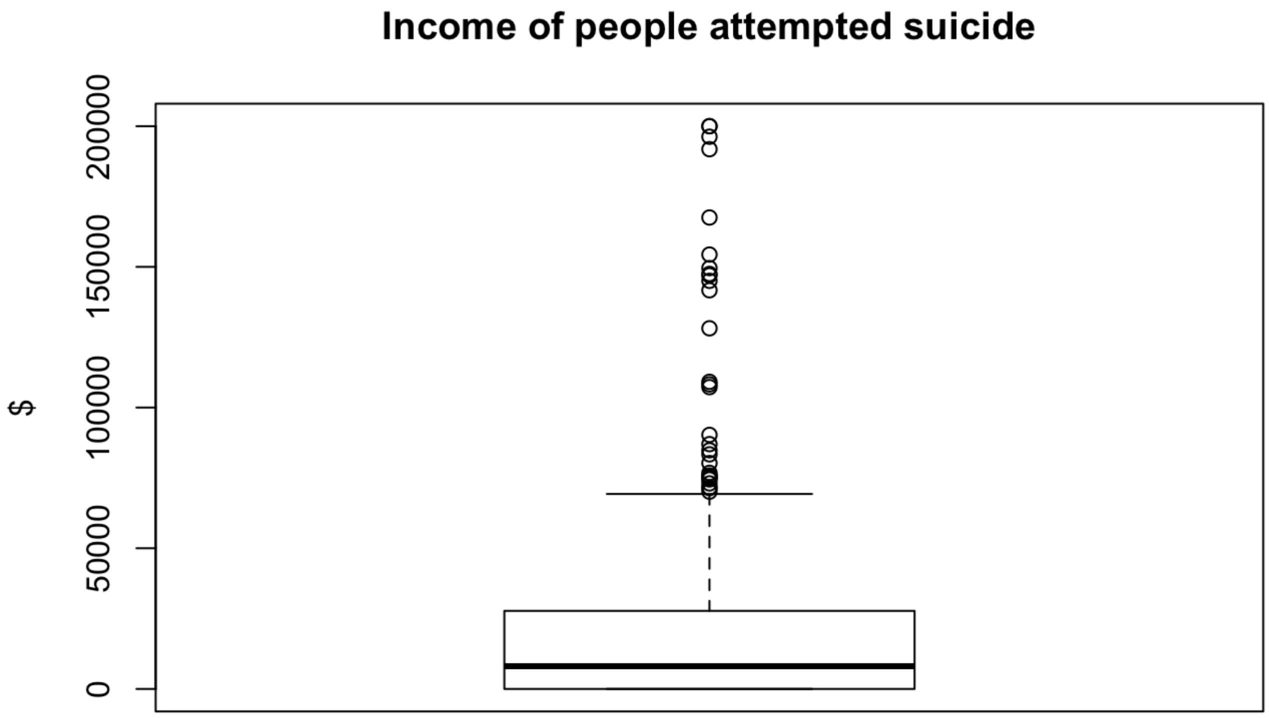
```
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

```
df = read.csv('rand_income.csv')
boxplot(df$income, main = 'Income of people attempted suicide', ylab = '$')
```



First, examine the boxplot of the income. The boxplot is not very help due to the high number of outliers. The box representing 50% of the data appears as a ligh segment rather than a box.

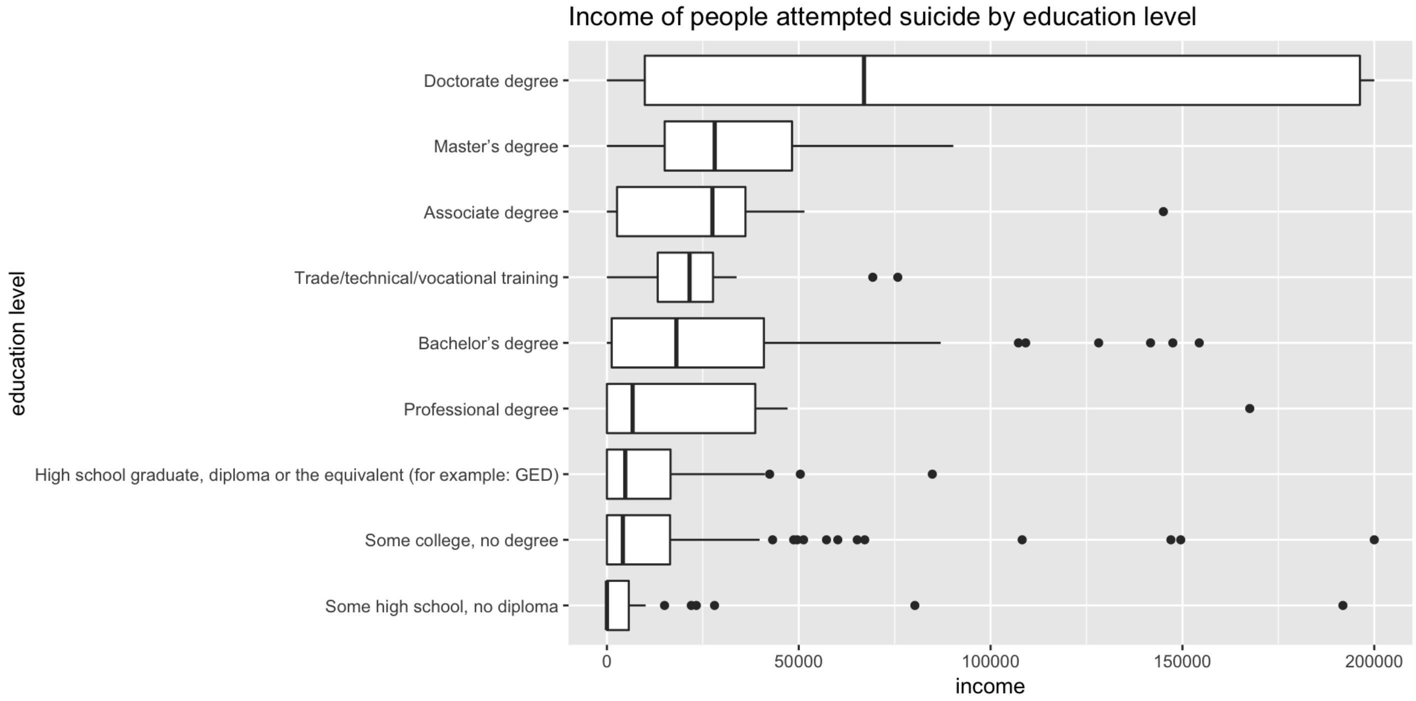
```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tid
yverse 1.2.1 ──
```

```
## ✔ tibble  1.4.2     ✔ purrr   0.2.5
## ✔ tidyr   0.8.1     ✔ stringr 1.3.1
## ✔ readr   1.1.1     ✔ forcats 0.3.0
```

```
## ── Conflicts ──────────────────────────────────────── tidyverse
_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
ggplot(df, aes(fct_reorder(edu_level, income, median), income)) +
  geom_boxplot() + xlab('education level') + coord_flip() +
  ggtitle('Income of people attempted suicide by education level')
```



Then, I wonder if faceting income by education level would give more insights.

People with Doctorate degree acquire the highest median income, then Master degree, Associate degree and Bachelor degree. Those with high school/no diploma have lowest income. This observation intuitively make sense, not only subjected to this suicide data but also in the larger population asa well because people with higher degree tend to find more profitable jobs.

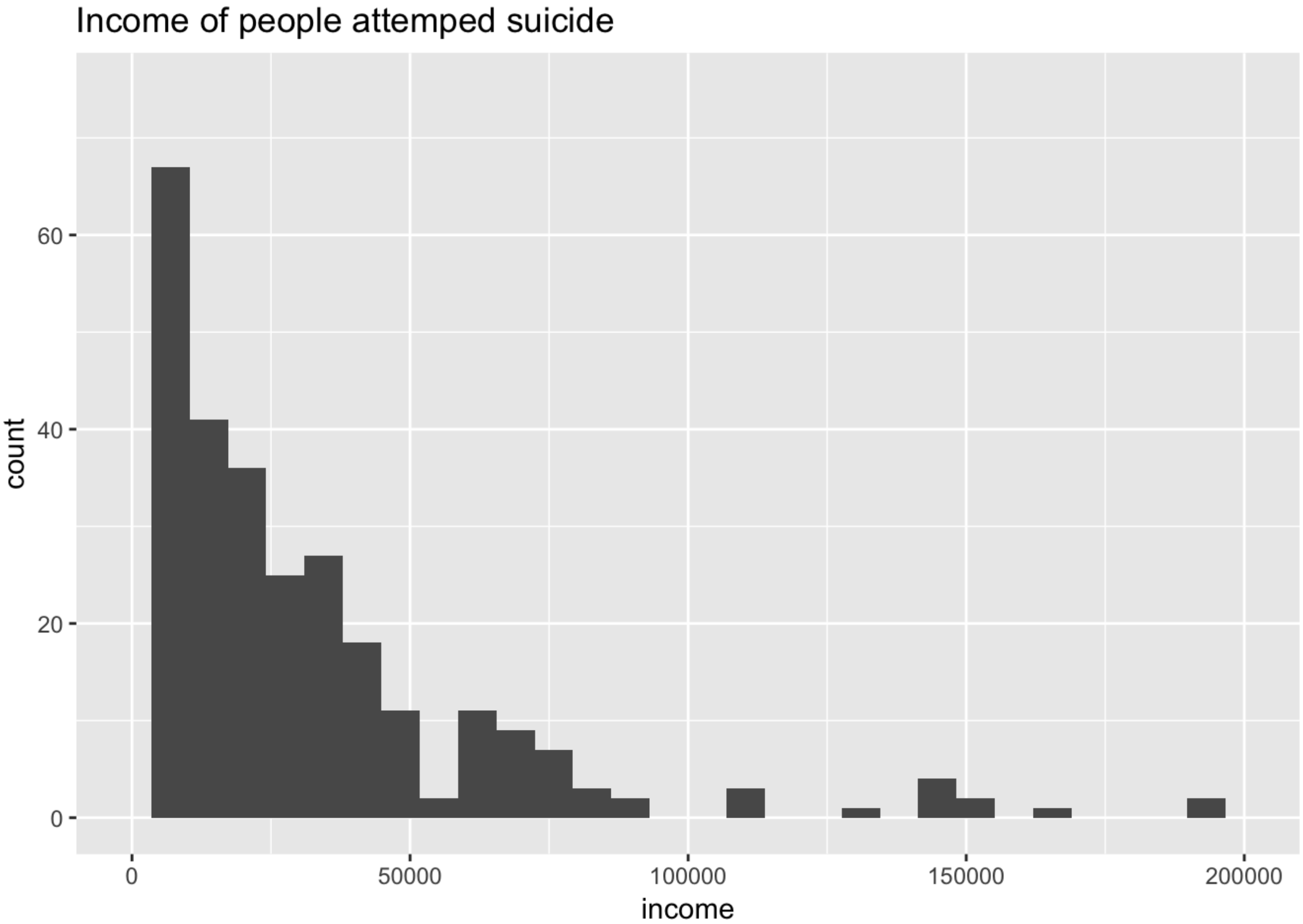
Moreover, Doctorate degree group are more spread out while other groups are more concentrated, which is a bit surprising. This tell us that the income of people with Ph.D degrees vary a lot, although relatively high in median.

There are also far more outliers in income from groups of Bachelor's degree, Some college degree and High school diploma. It indicates that the income of people with less sophisticated education have fat heads/fat tails

```
ggplot(df, aes(income)) + geom_histogram() +
  ggtitle('Income of people attemped suicide') + ylim(0, 75) + xlim(0,200000)
```

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

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

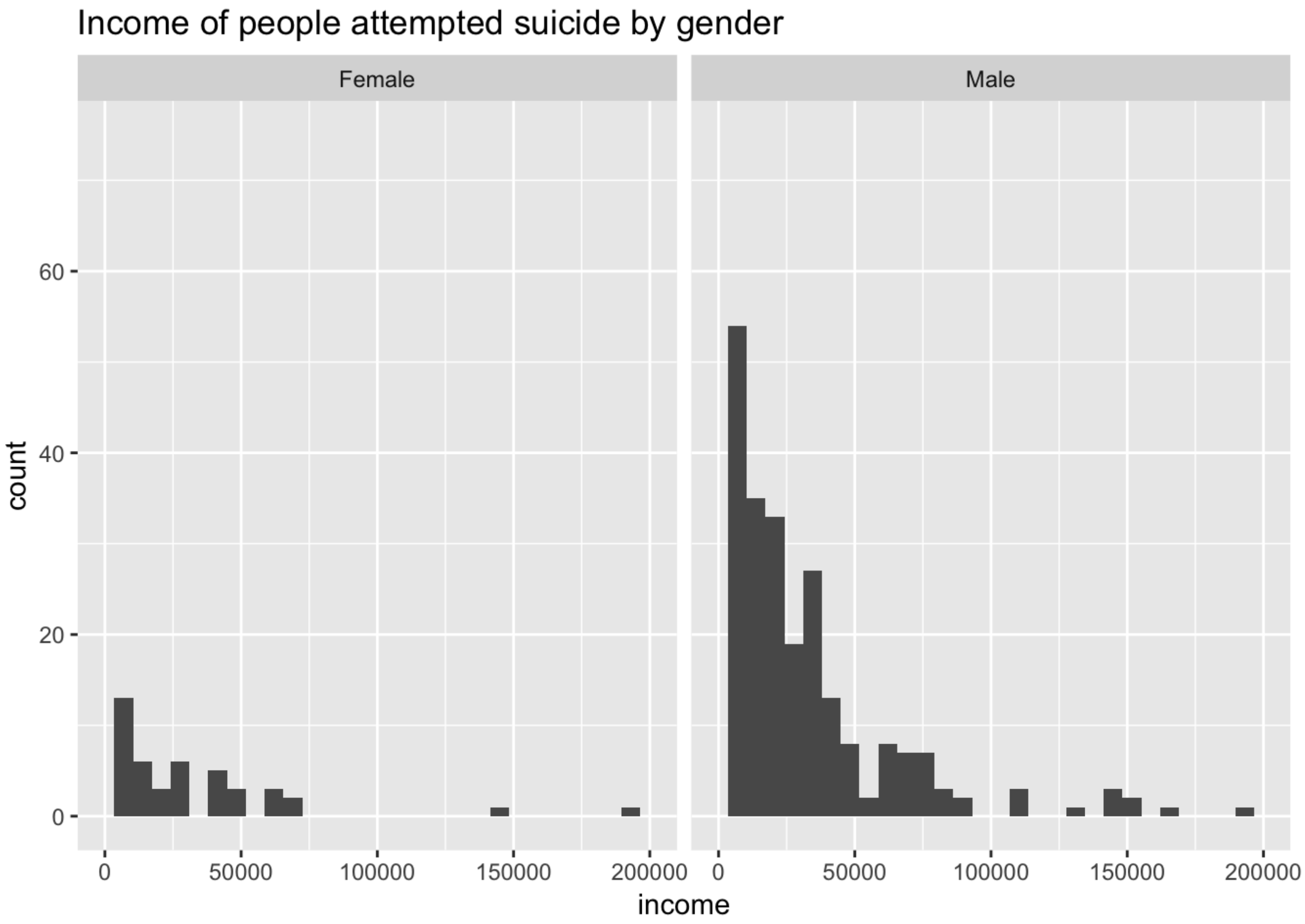


As shown by the histogram: 1. More people with lower income committed suicide than those with higher income. 2. There exist outliers: people with exceptionally high income committed suicide. 3. Among people who committed suicide, large number of them have income of $0, i.e. do not have income.

```
df <- df %>% filter(df$gender == 'Male' | df$gender == 'Female')
ggplot(df, aes(income)) + geom_histogram() +
  ggtitle('Income of people attempted suicide by gender') + ylim(0, 75) + xlim(0,200000)
 + facet_wrap(~gender)
```

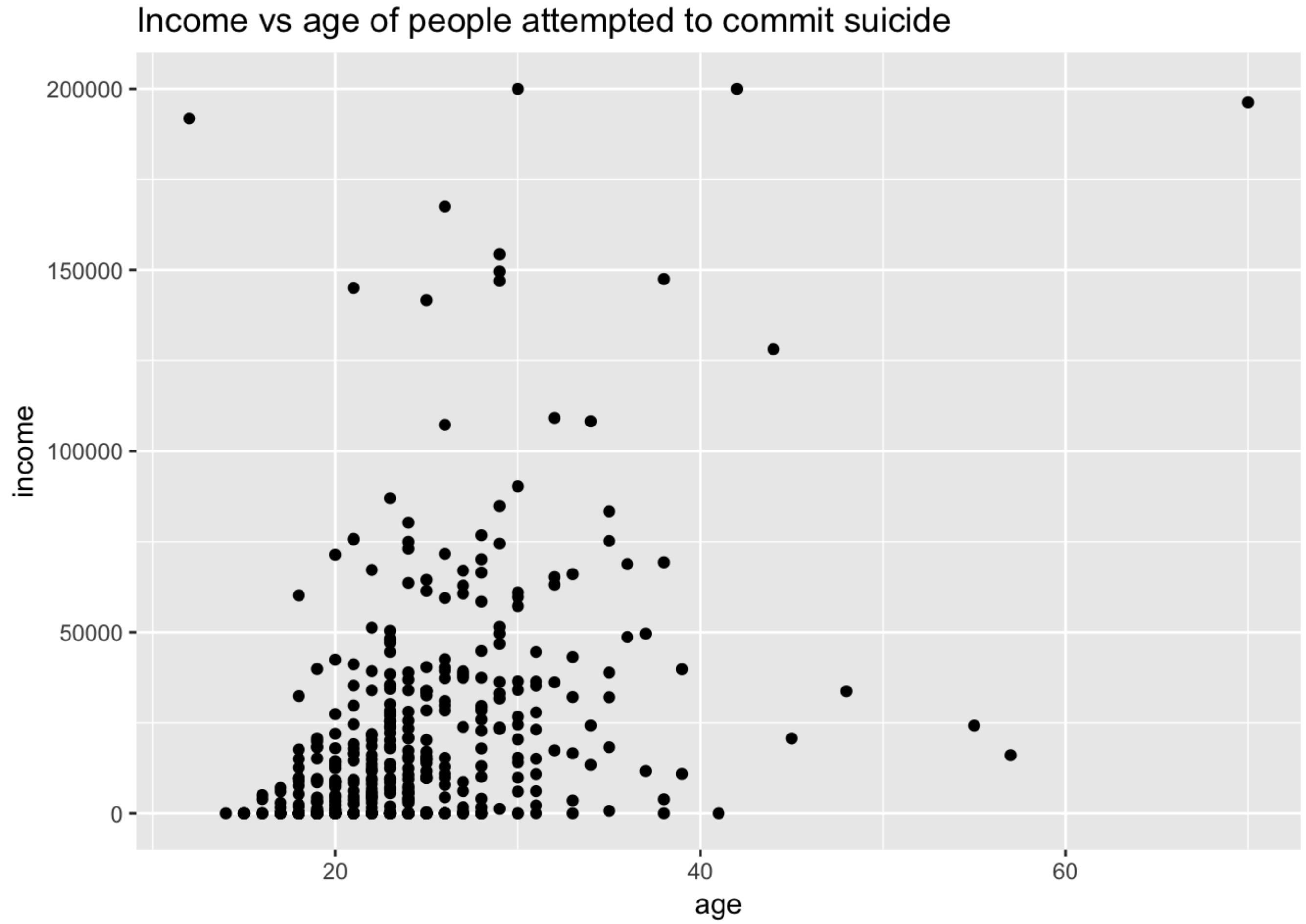
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



Facet by gender shows that Male’s income is representive because there are far more males than females in this dataset. As the data sizes of two groups are not comparable, there is not enough information to furthur infer the contrast between two distributions.

```
ggplot(df, aes(x = age, y = income)) + geom_point() +
  ggtitle('Income vs age of people attempted to commit suicide')
```



I wonder if there's relationship between age and income among people who attempted to commit suicide. And observe the following:

1. There are people with no income, regardless of their ages.
2. There are people with high income, regardless of their ages.
3. There are fewer people above age 40, large number of people between 20-40 years old.
4. There are fewer people with higher income who attempt to commit suicide.
5. As indicated, there is no strong relationship between age and income among this group, which is counter-intuitive.