Exploratory Data Analysis Project

Carolina Garcia

2024-03-01

```
knitr::opts_chunk$set(fig.width=6, fig.height=4)
knitr::opts_chunk$set(warning=FALSE, message=FALSE)
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
```

```
library(RColorBrewer)
depression <- read.table("../data/depression.txt", header=TRUE, sep="\t")
```

Introduction

The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

Univariate Exploration

Lets first look at education among those who were interviewed

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

```
r
ggplot(depression,
aes(x=education))
+
theme_minimal()
+
geom_bar(aes(y
=
..count..))
+
ggtitle("Increasing
levels of
Education
vs. People
Surveyed")
+
geom_text(aes(y=..count..
+ 10,
label=..count..),
stat='count',
size = 3)
```

Introduction The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

```
## Error in
`geom_bar()`:
## !
Problem
while
computing
aesthetics.
## i Error
occurred in
the 1st
layer. ##
Caused by
error: ## !
object
'education'
not found
```

Here we can see the range of people and the increasing levels of education. We can see that more than half of the people surveyed have gotten a high school diploma or higher.

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

Next lets look at marital factors

```
r
depression$MARITAL
<-
factor(depression$MARITAL,
labels =
c("Widowed",
"Divorced",
"Married",
"Never
Married",
"Separated"))
table(depression$MARITAL)
```

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

```
## ##
Widowed
Divorced
Married
Never
Married
Separated
##
73
127
43
13
38
r
(43/294)*100
## [1]
14.62585
r
(251/294)*100
```

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

```
## [1] 85.37415
```

Here a table was used to seperate the 294 people into 5 categories that explain their marital status. Using the table we can then calculate the percent of married people, 14.63%, and those who fall within the rest of the categories, 85.37%.

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

Next lets move on to employment status

```
r
employment
<-
factor(depression$EMPLOY,
labels =
c("Full
Time",
"Part
Time",
"Unemployed",
"Retired",
"House
person",
"In
school",
"Other"))
table(depression$EMPLOY)
## ##   1
2   3   4
5   6   7
## 167  42
14  38  27
2   4
```

Introduction
The data set to be analyzed is in regards to depression. The data set comes from Los Angeles County and it inculdes 294 observations of people who where interviewed about depression. The depression data set will be used to see what will affect a person's mental health.

```
r
sum(14,38,27,2,4)
## [1] 85
r
(85/294)*100
## [1]
28.91156
r
(209/294)*100
## [1]
71.08844
```

Here we see that about 29% of the people do not work, while about 71% of people have a job.

Bivariate Exploration

```
table(depression$MARITAL, depression$EMPLOY)
```

```
##
```

```
##                   1  2  3  4  5  6  7
##   Widowed        49 13  5  3  1  2  0
##   Divorced       72 19  4 12 17  0  3
##   Married        30  4  3  3  2  0  1
##   Never Married   7  2  1  2  1  0  0
##   Separated       9  4  1 18  6  0  0
```

When comparing marital staus and employment status we see that diverced people have the largest amount in more than half the catergories and as usual anyone who is not married takes up more than half the values on the table. We can make an assumption that people who are not married and work full time are going to be depressed.

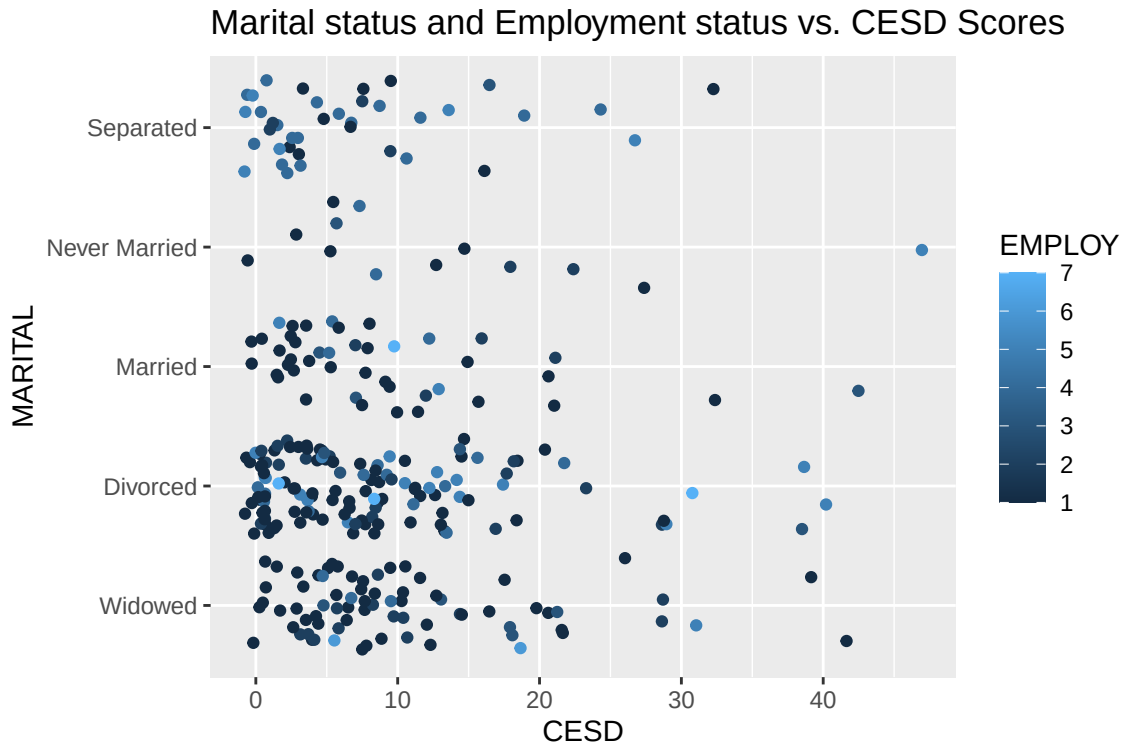
Next we will compare marital status and education level vs. CESD Scores

```
ggplot(depression, aes(x = CESD, y = MARITAL , color = education)) + geom_jitter(width = 0.8) +
ggtitle("Marital Status and Education Level vs. CESD Scores")
```

```
## Error in `geom_jitter()`:
## ! Problem while computing aesthetics.
## i Error occurred in the 1st layer.
## Caused by error:
## ! object 'education' not found
```

This plot is data heavy within the divorced and widowed sections, so we can assume those who fall within these categories have daily thoughts of depression

```
ggplot(depression, aes(x = CESD, y = MARITAL , color = EMPLOY)) + geom_jitter(width = 0.8) +
ggtitle("Marital status and Employment status vs. CESD Scores")
```



This plot is data heavy within the divorced and widowed sections once again, so we can assume those who fall within these categories have daily thoughts of depression

And finally we can look at marital status and employment status vs. CESD scores

```
ggplot(depression, aes(x = CESD, y = education , color = EMPLOY)) + geom_jitter(width = 0.8) +
ggtitle("Marital status and Employment status vs. CESD Scores")
```

```
## Error in `geom_jitter()`:
## ! Problem while computing aesthetics.
## i Error occurred in the 1st layer.
## Caused by error:
## ! object 'education' not found
```

We can observe that those with higher levels of education are less likely to have daily thoughts of depression

---

Conclusion The univariate section let us explore the the percentage of education, marital status, and employment rates among the people surveyed. This section let us see that most of the people who have a hischool education or higher, aren't married, and have a job make up most of the depressed population surveyed. Among our bivariate scatter plots we can see a commonality among those that are Married. This group of people will normally rank the lowest with depression or daily thoughts of depression. This is also supported by our univariate findings. The hypothesis is that those who are single, lack employment, and have had little education will have the highest depression rates was incorrect and is the opposite for 2/3 of those cases.