Quiz_1

2025-02-07

Section 1: Blitz Questions (You can write answers as a comment in R)

- What does the interquartile range (IQR) represent?

It represent the data in the (Q_1, Q_3) range. In simpler terms it represents the middle 50% of the sorted data.

- How can you tell if data is skewed by looking at a histogram?

If more of the data is gathered either at the left or right we can say that it is skewed.

- Imagine you have a data named my_data that contains one numeric (col1) and one categorical column (col2). Look at the script bellow, and describe how errors can be fixed ? (You can just write fixed version without description)

ggplot(my_data=data, aes(x="col1", fill="col2")) + geom_histogram

Fixed: ggplot(data, aes(x=col1, fill=col2)) + geom_histogram()

Section 2: Create environment and run this script.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

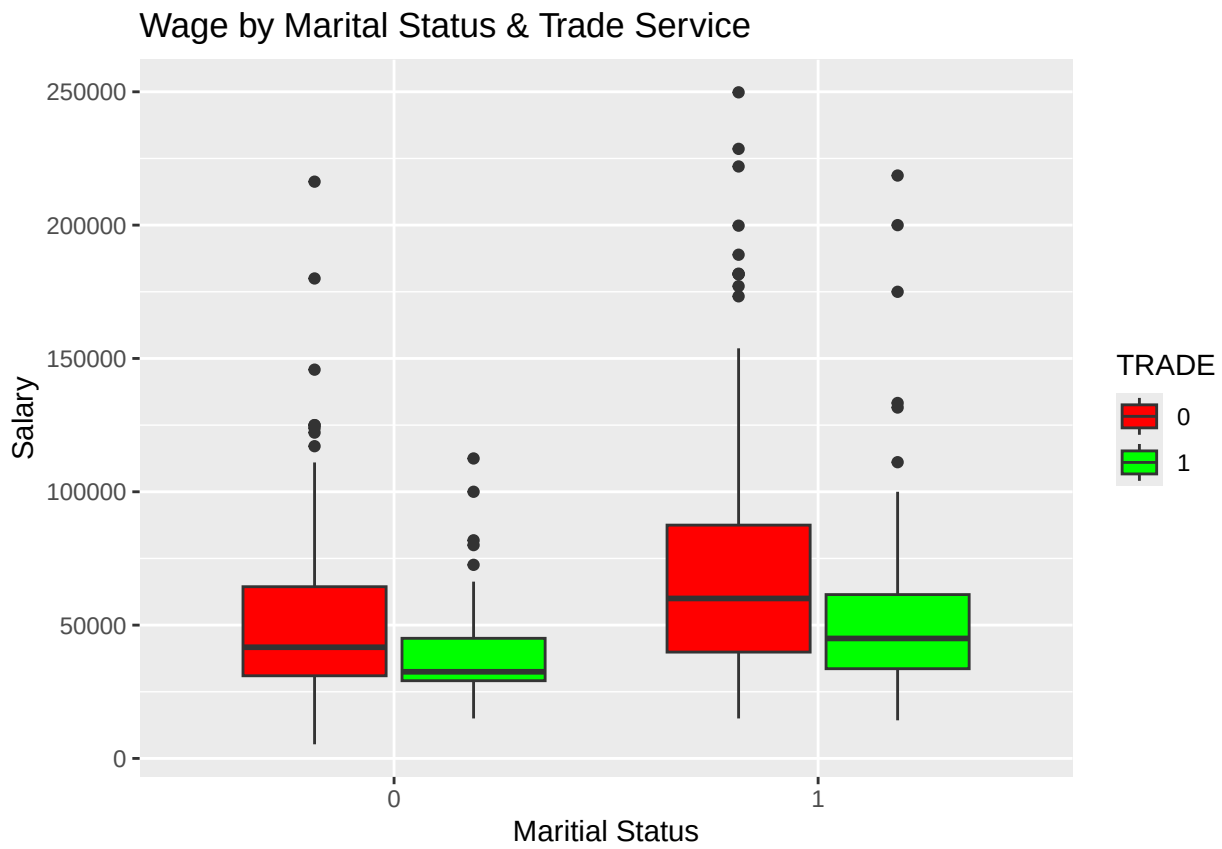
```
##
## Attaching package: 'dplyr'
```

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

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

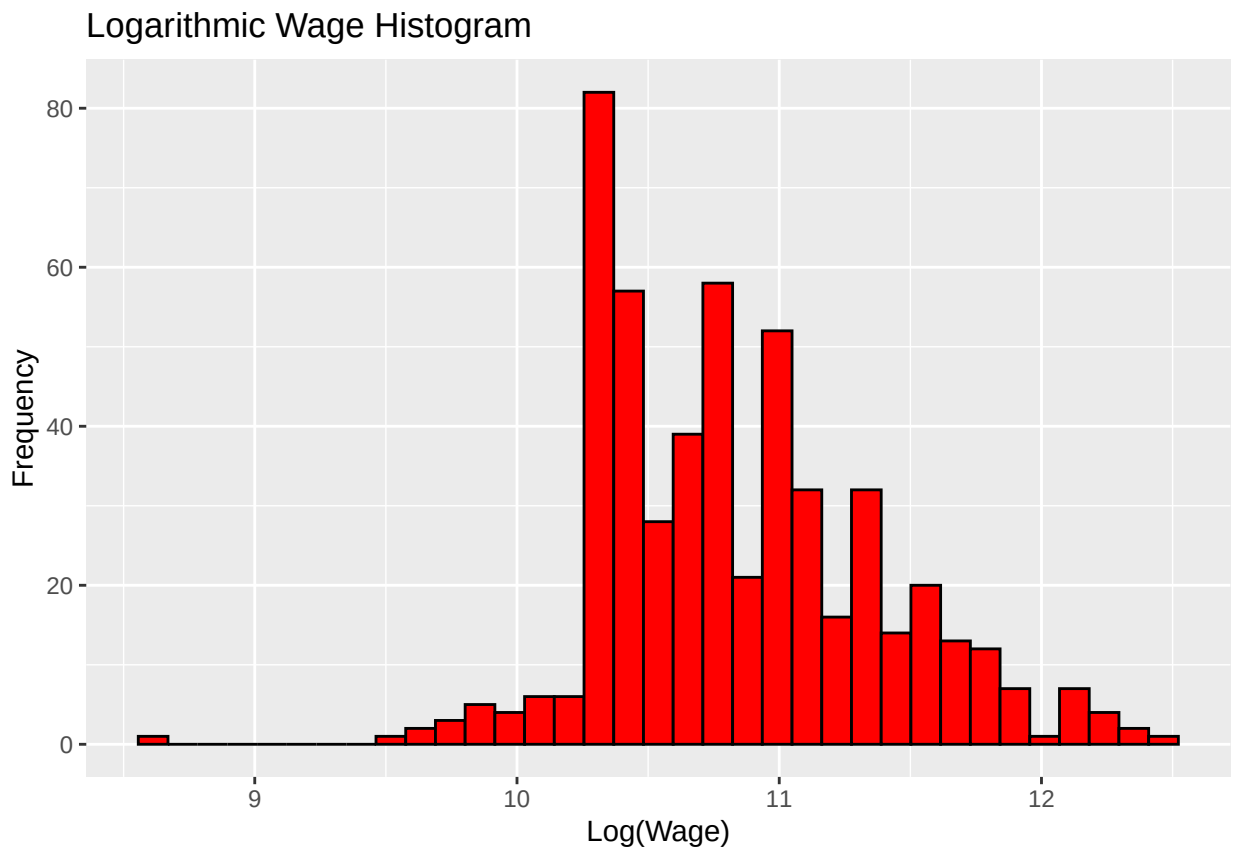
```
library(ggplot2)
wage_data <- read.csv("wage_data.csv", sep=";")
wage_data$LWAGE <- log(wage_data$WAGE)
categorical_cols <- c("MARRIED", "NONWHITE", "NORTHCEN", "PROFSERV",
"SOUTH", "TRADE", "WEST", "FEMALE")
wage_data[categorical_cols] <- lapply(wage_data[categorical_cols], as.factor)


ggplot(data=wage_data, aes(x=MARRIED, y=WAGE, fill=TRADE)) + labs(title="Wage by Marital Status & Trade
```



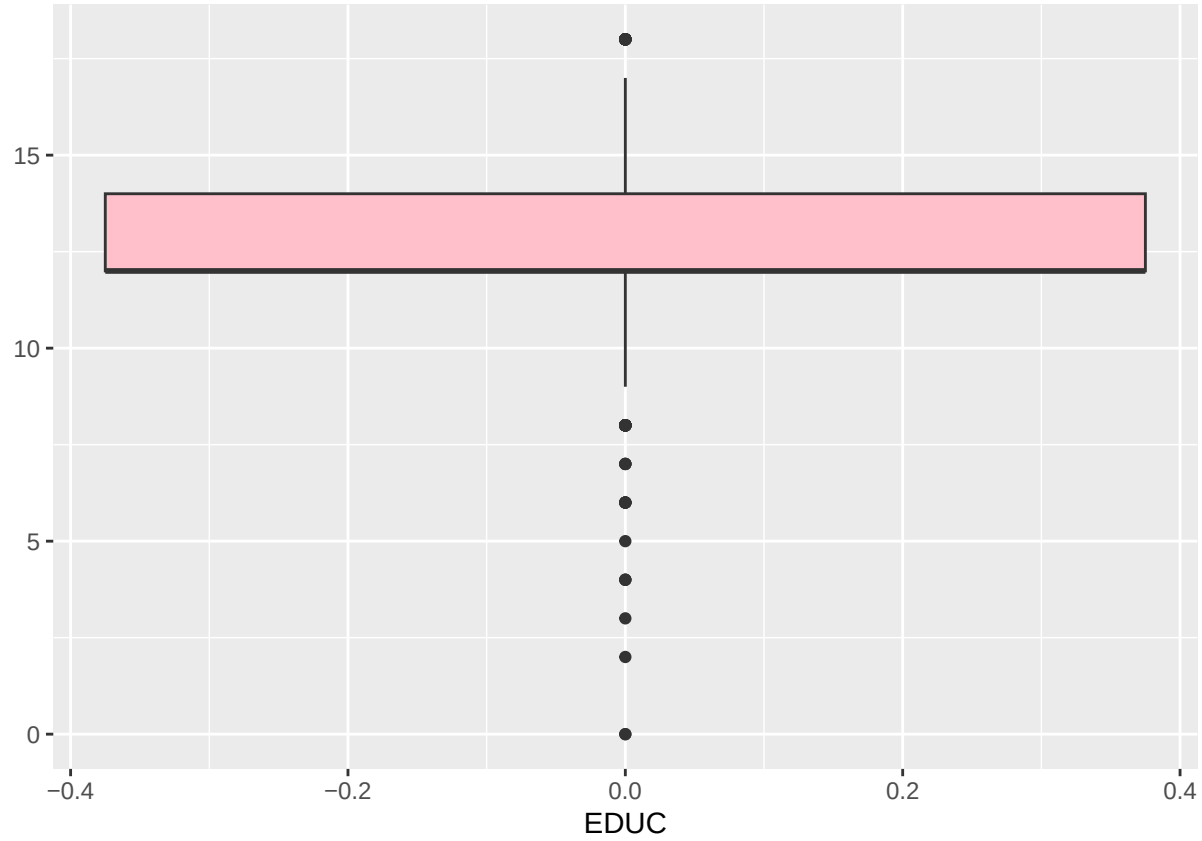
Task 2.
- Create Histogram for LWAGE column
- Add title to your graph
- Fill the graph with any color that you want
- Change bin border colors to some other color
- Change the X and Y axis names
- Set bins equal to 35

```
ggplot(data=wage_data, aes(x=LWAGE)) + labs(title="Logarithmic Wage Histogram", x ="Log(Wage)", y = "Fr
```



Task 3: Recreate following graph using years of education column and answer question bellow. Note:

```
ggplot(data=wage_data, aes(x=EDUC)) + coord_flip() + labs(x = "", y ="EDUC") + geom_boxplot(fill="pink"
```



Question 1: How many outliers are there in the graph?

At least 9. Since they can coincide we can not say definitely but we can at least say that there are 9 or more outliers.

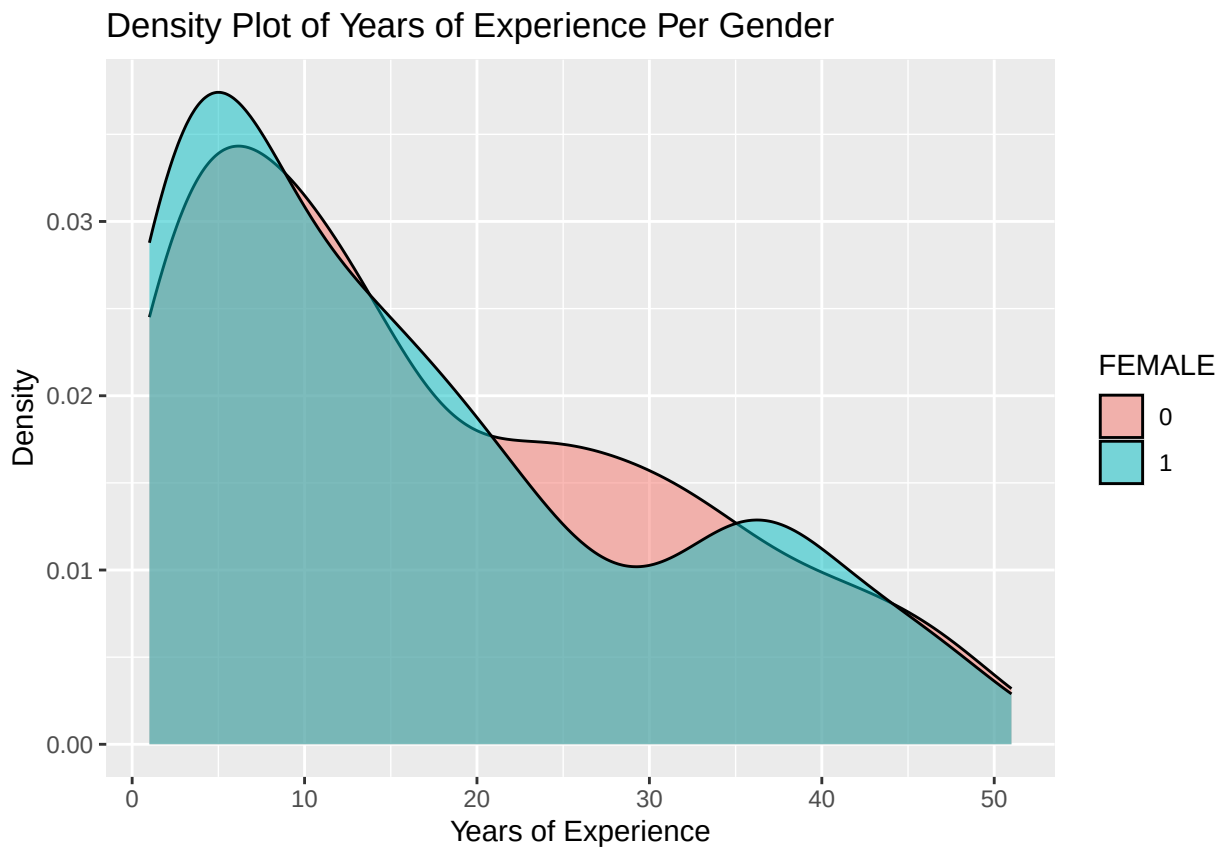
Question 2: Why in this graph we don't see median?

The median is at the very bottom we don't see it because it coincides with Q_1.

Task 4: Create density plot for years of experience variable for each gender and add transparency to your graph. Is there any difference between the distributions?

Yes from the graph we can see that Females are generally more inclined to have experience from 0-10 years as apposed to Males, and Males are more inclined to have experience in the range 20-35 years. Afterwards it about equally dense.

```
ggplot(data=wage_data, aes(x=EXPER, fill=FEMALE)) + labs(title = "Density Plot of Years of Experience P
```



Section 3: (Optional for bonus point) • How many observations are in this graph?

4+4+6+7+6+19+12+11+7+12+7+2+2+1=100

• How many bins are in this graph?

15