Math 170S Homework #2: Exploratory Data Analysis

Damien Ha

2023-01-25

Contents

Loading Data and Packages **1**

Univariate Analysis **2**
- Sex . . . 2
- Age . . . 2
- Salary . . . 5
- Expense . . . 7

Bivariate/Multivariate Analysis **10**
- Expense vs Salary . . . 10
- Salary vs Age . . . 11
- Expense vs Age . . . 13
- Scatterplot Matrix of Age, Salary, and Expense . . . 14

Loading Data and Packages

```
suppressWarnings(library(ggplot2))
df <- read.csv("HW2data.csv")
head(df)
```

```
##   ID    Sex Age Salary Expense
## 1 12   Male  22   2311    1050
## 2 13   Male  24   3231    1265
## 3 14   Male  27   2423    1109
## 4 15   Male  19   3343    1511
## 5 16 Female  20   2535    1147
## 6 17 Female  24   3455    1564
```

```
tail(df)
```

```
##    ID    Sex Age Salary Expense
## 46 39   Male  24   4877    2190
## 47 39 Female  28   4069    1830
## 48 39   Male  30   4989    2246
## 49 39 Female  33   4181    1884
## 50 39   Male  25   5101    2292
## 51 39   Male  26   4293    1929
```

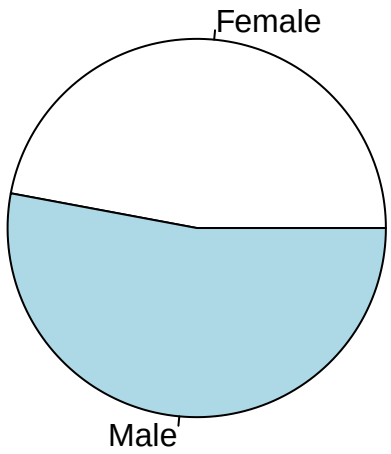
Here is a function for calculating the mode

```r
getmode <- function(v) {
    uniqv <- unique(v)
    uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Univariate Analysis

Sex

```r
pie(table(df$Sex), main = "Sex of Indidividuals in Data")
```



There is a near even split between male and female in this data, but with slightly more male

Age

```r
summary(df$Age)
```

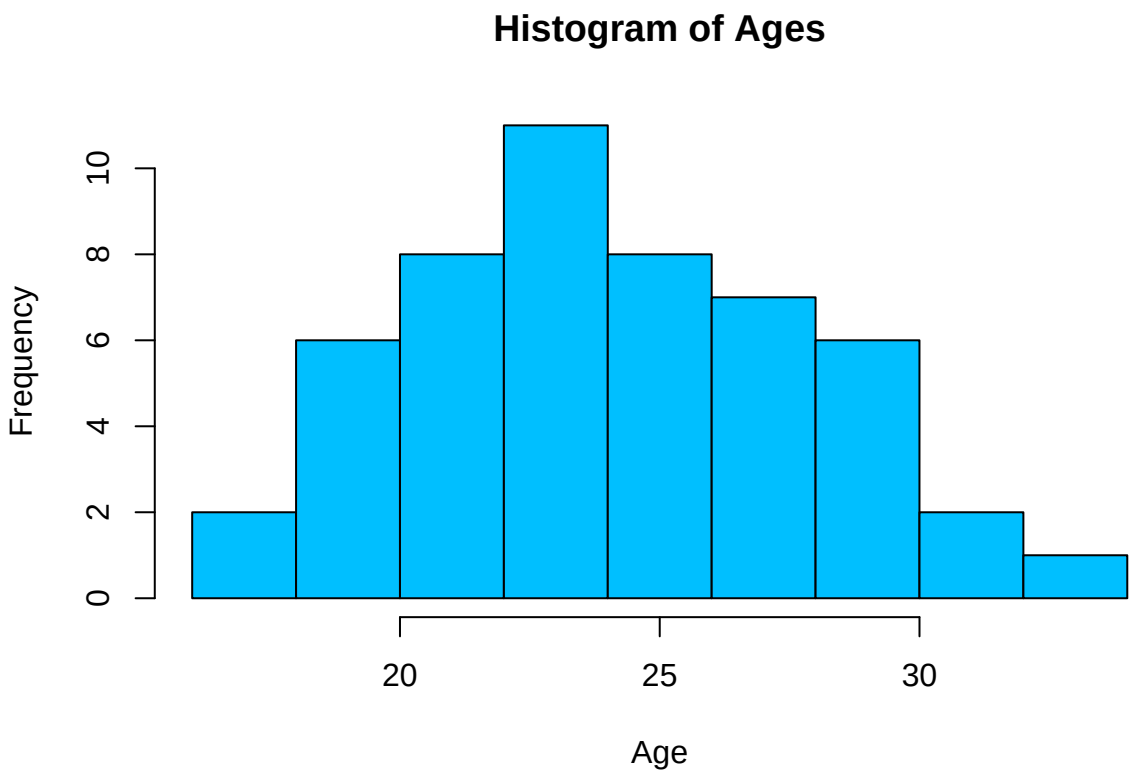
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   22.00   24.00   24.51   27.00   33.00
```

```r
getmode(df$Age)
```

```
## [1] 24
```

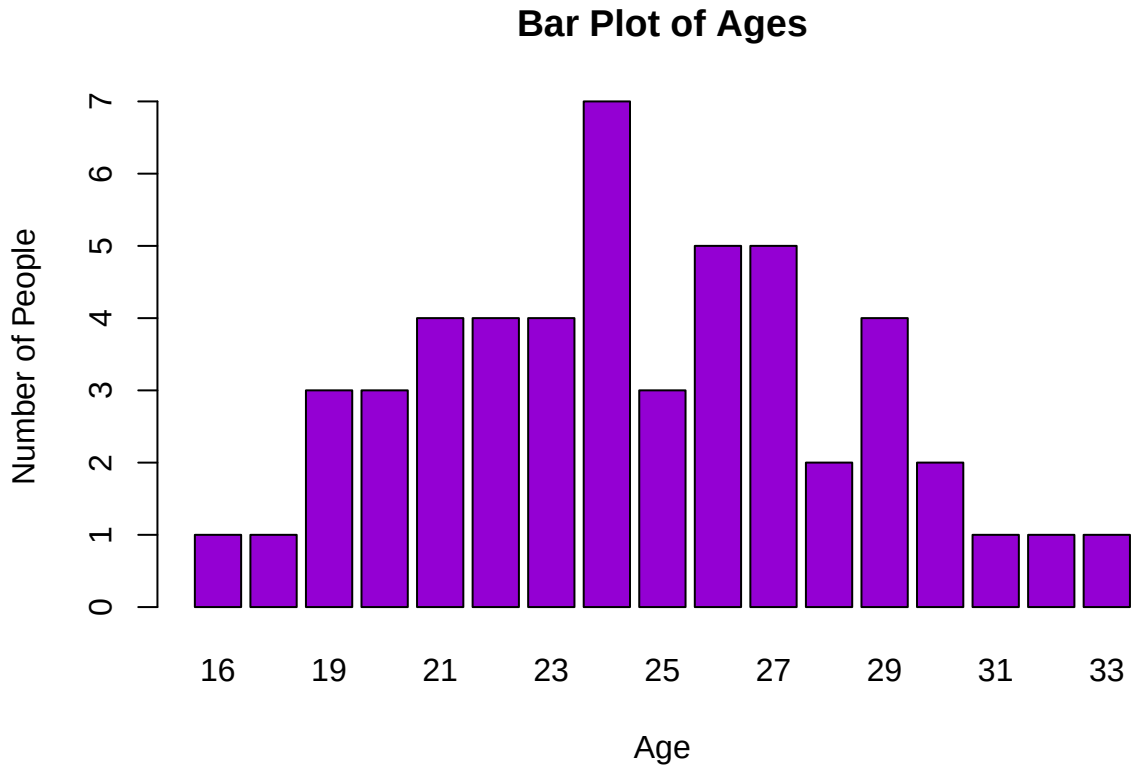
So the mean age is 24.51, the median age is 24, and the mode/most common age is also 24.

```r
hist(df$Age, xlab = "Age", main = "Histogram of Ages", col = "deepskyblue")
```

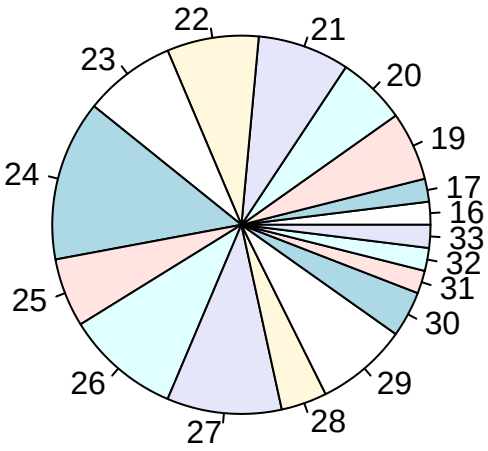


Age looks to be distributed fairly normally

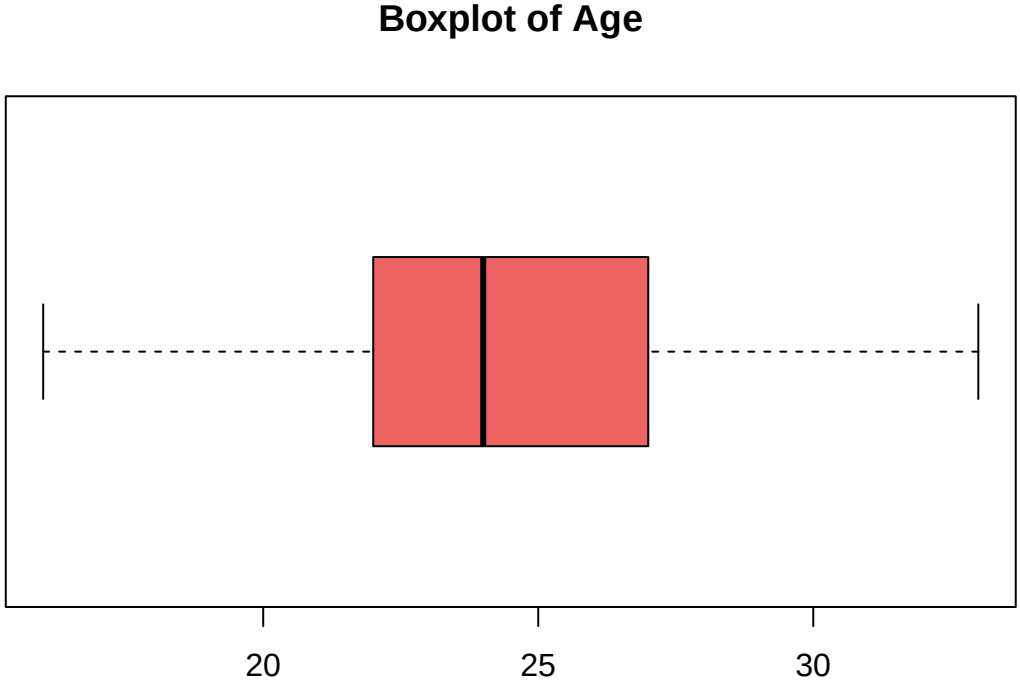
```
barplot(table(df$Age), xlab = "Age", ylab = "Number of People", main = "Bar Plot of Ages", col = "darkv
```




```
pie(table(df$Age), main = "Pie Chart of Ages")
```

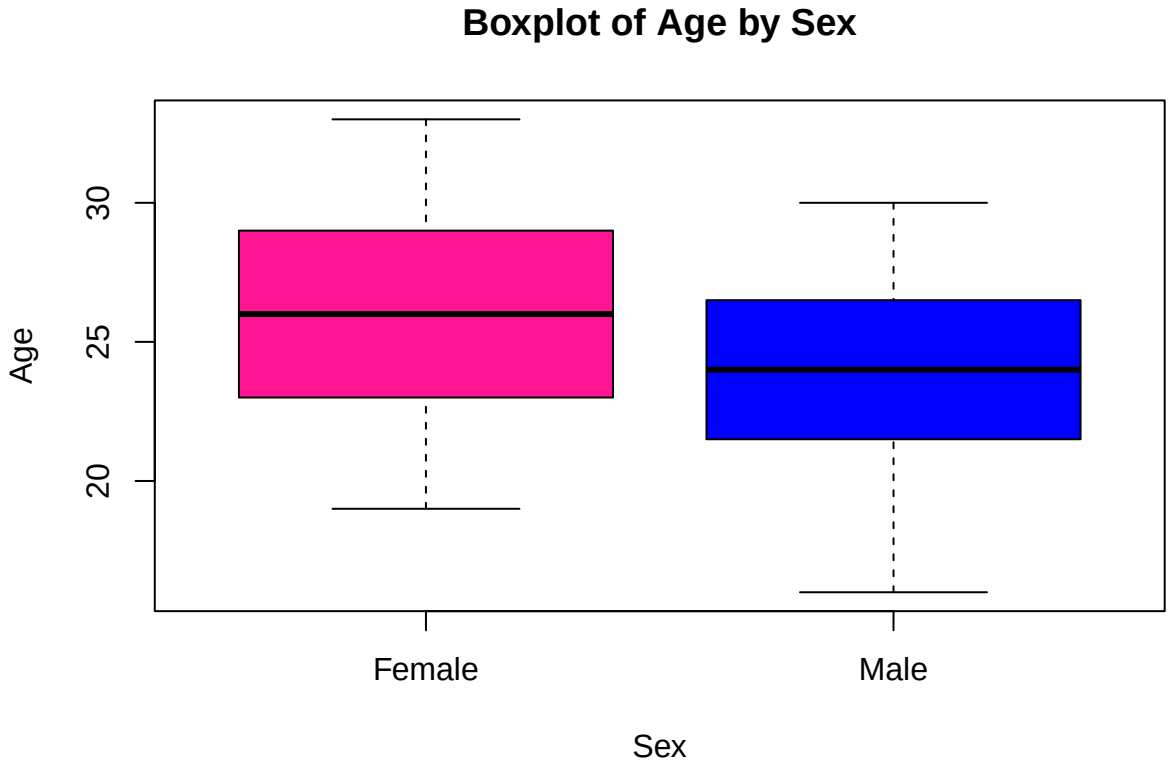



```
boxplot(df$Age, col="indianred2", plot=T, horizontal=T, main="Boxplot of Age")
```



There are fewer people at the oldest and youngest ends of the spectrum, i.e. fewer 16 and 17 year olds and 32 or 33 year olds than people in their 20s

```
boxplot(df$Age ~ df$Sex, col=c("deeppink","blue"), xlab="Sex", ylab="Age", main="Boxplot of Age by Sex")
```



There seem to be more older female indidviduals in the data compared to males

Salary

```
summary(df$Salary)
```

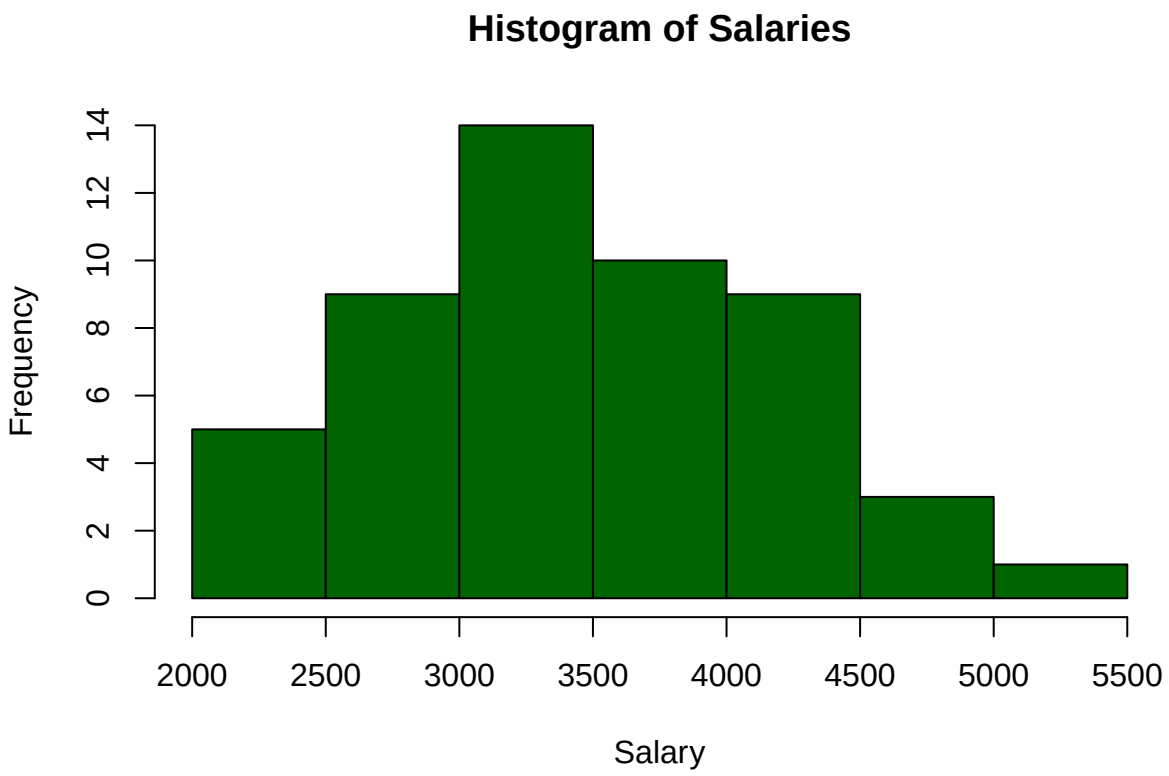
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2191    2991    3447    3472    4002    5101
```

```
getmode(df$Salary)
```

```
## [1] 2311
```

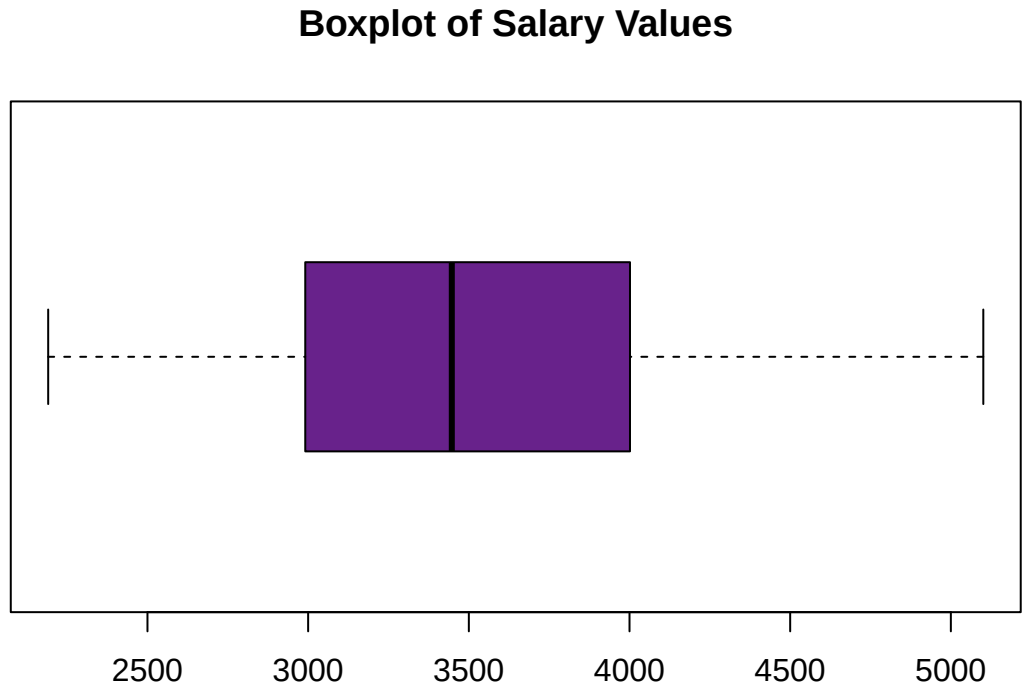
So the average salary is 3472, the median salary is 3447, and the most common salary is 2311. There is a large jump from the mode to the mean and median; larger values/possible outliers may have some influence on the mean. We can see that the max value is much higher than the 3rd quartile

```
hist(df$Salary, xlab = "Salary", main = "Histogram of Salaries", col = "darkgreen")
```



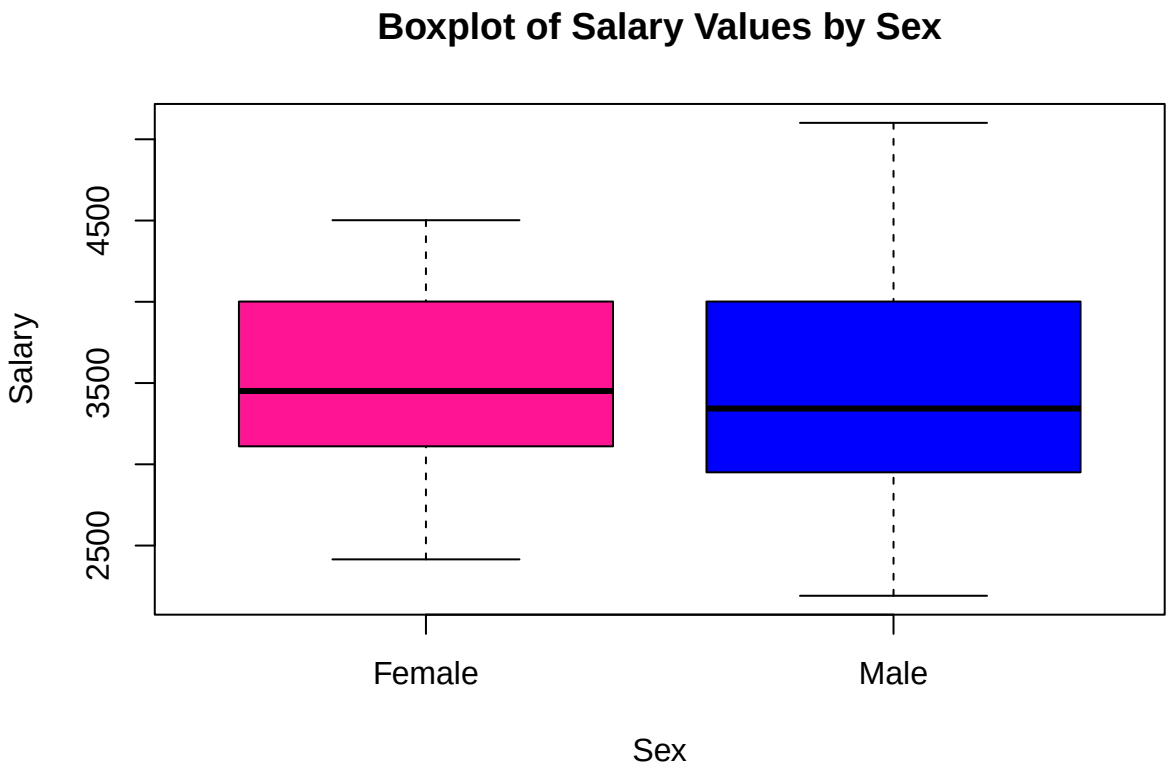
Salary looks slightly right skewed, which supports what we saw in our previous observation.

```
boxplot(df$Salary, col="darkorchid4", plot=T, horizontal=T, main="Boxplot of Salary Values")
```



Now let's look at this by sex

```
boxplot(df$Salary ~ df$Sex, col=c("deeppink","blue"), xlab="Sex", ylab="Salary", main="Boxplot of Salary
```



Males in this data have a wider range of salary than females

Expense

```
summary(df$Expense)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     990    1324    1548    1559    1794    2292
```

```
getmode(df$Expense)
```

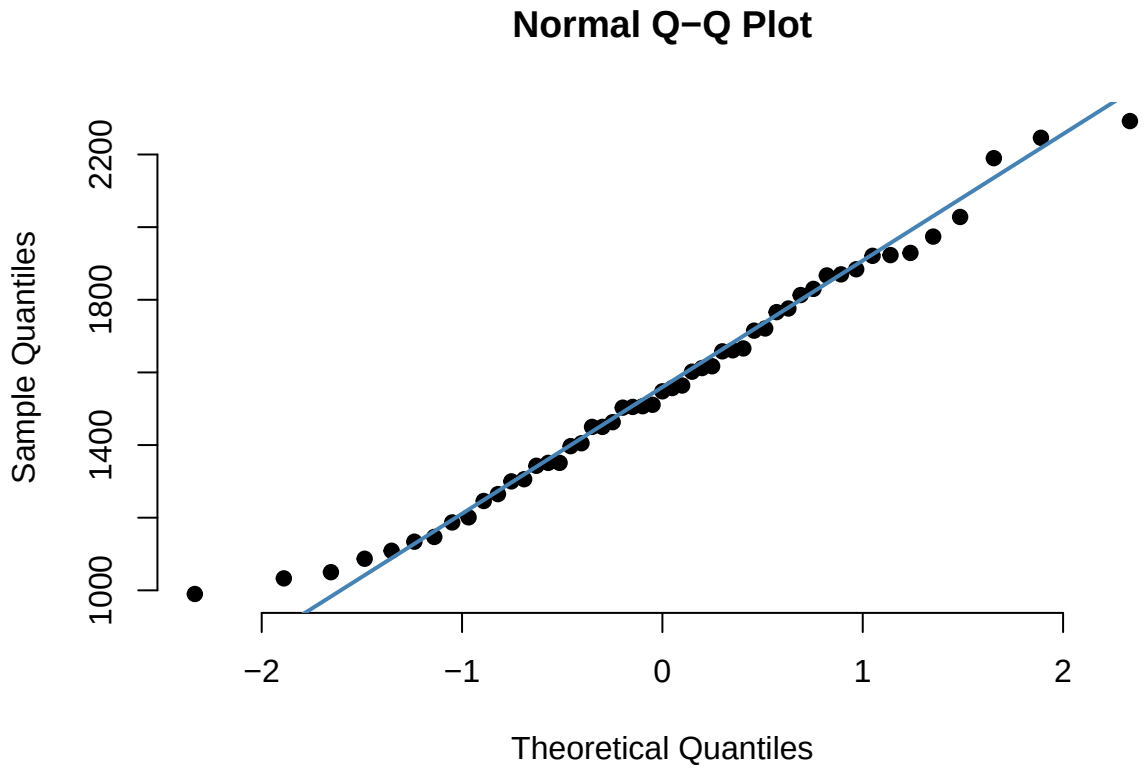
```
## [1] 1351
```

So the mean is 1559, the median is 1548, and the mode is 1351. Again, large expense values may be increasing the mean

```
hist(df$Expense, xlab = "Expense", main = "Histogram of Expense Values", col="darkgoldenrod1")
```

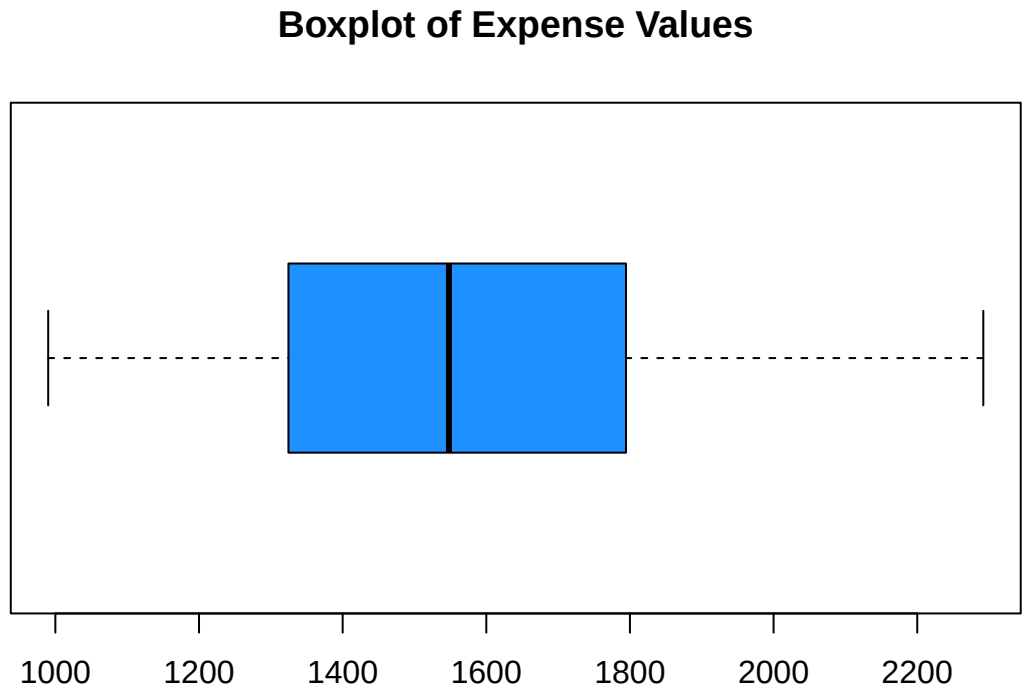
Expense is not right skewed like salary, but the larger values around 2000 do seem to lie a bit more outside the rest of the data

```
qqnorm(df$Expense, pch = 19, frame = FALSE)
qqline(df$Expense, col = "steelblue", lwd = 2)
```

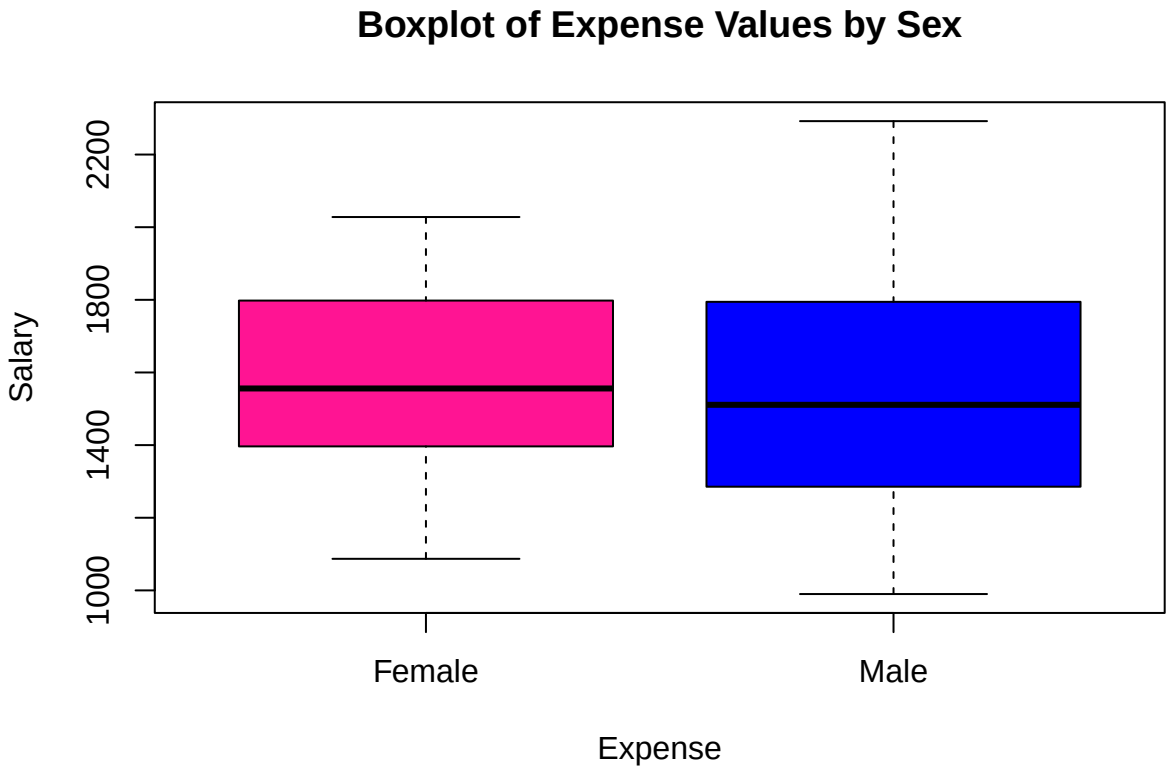


From the qqplot, expense is fairly normally distributed. The points on the plot seem to fall on a fairly straight line and don't seem to have some other pattern that would imply skewedness

```
boxplot(df$Expense, col="dodgerblue", plot=T, horizontal=T, main="Boxplot of Expense Values")
```




```
boxplot(df$Expense ~ df$Sex, col=c("deeppink","blue"), xlab="Expense", ylab="Salary", main="Boxplot of
```



Males in this data also have a wider range of expenses than females

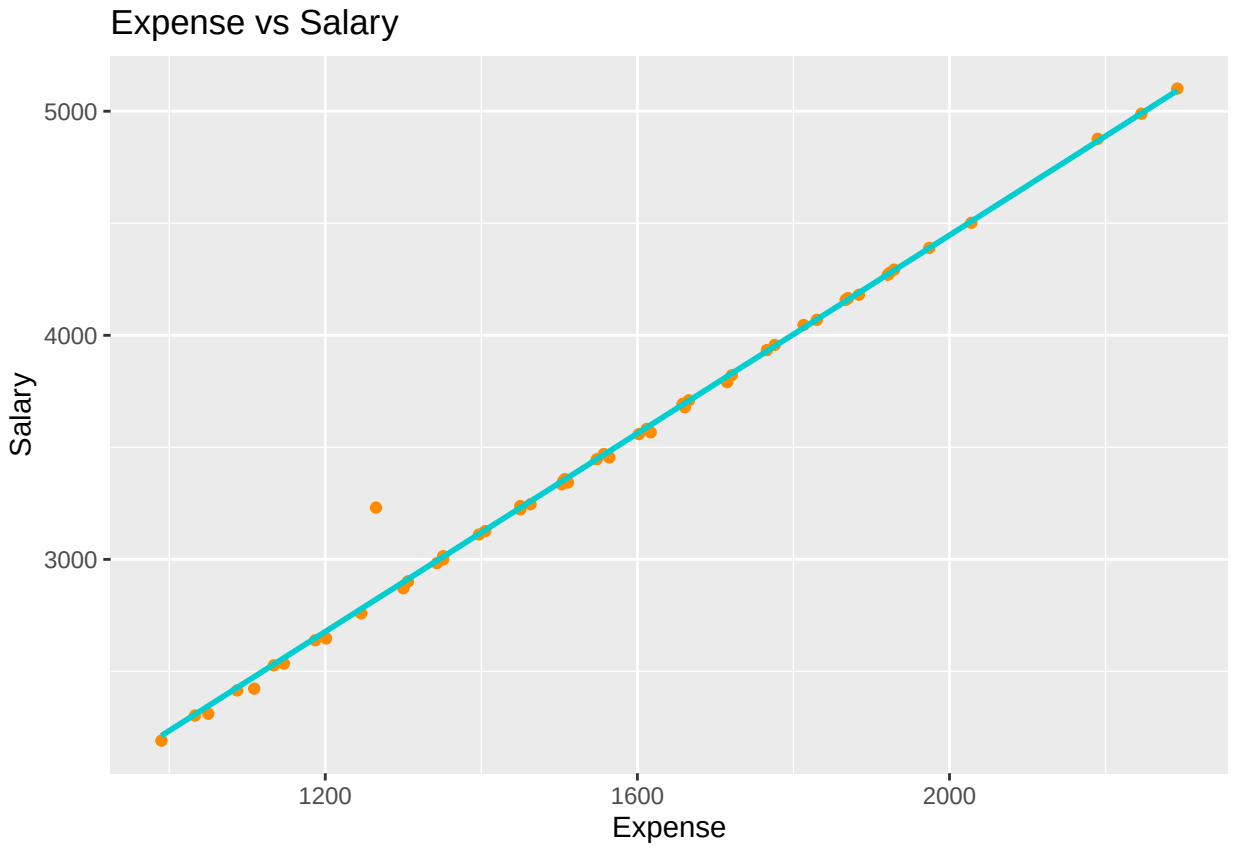
Bivariate/Multivariate Analysis

Expense vs Salary

```
ggplot(df, aes(x=Expense, y=Salary)) + geom_point(colour="darkorange") + stat_smooth(method = "lm", col
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(df, aes(x=Expense, y=Salary)) + geom_point(aes(color = factor(Age))) + labs(colour = "Age", titl
```


```
cor(df$Expense, df$Salary)
```

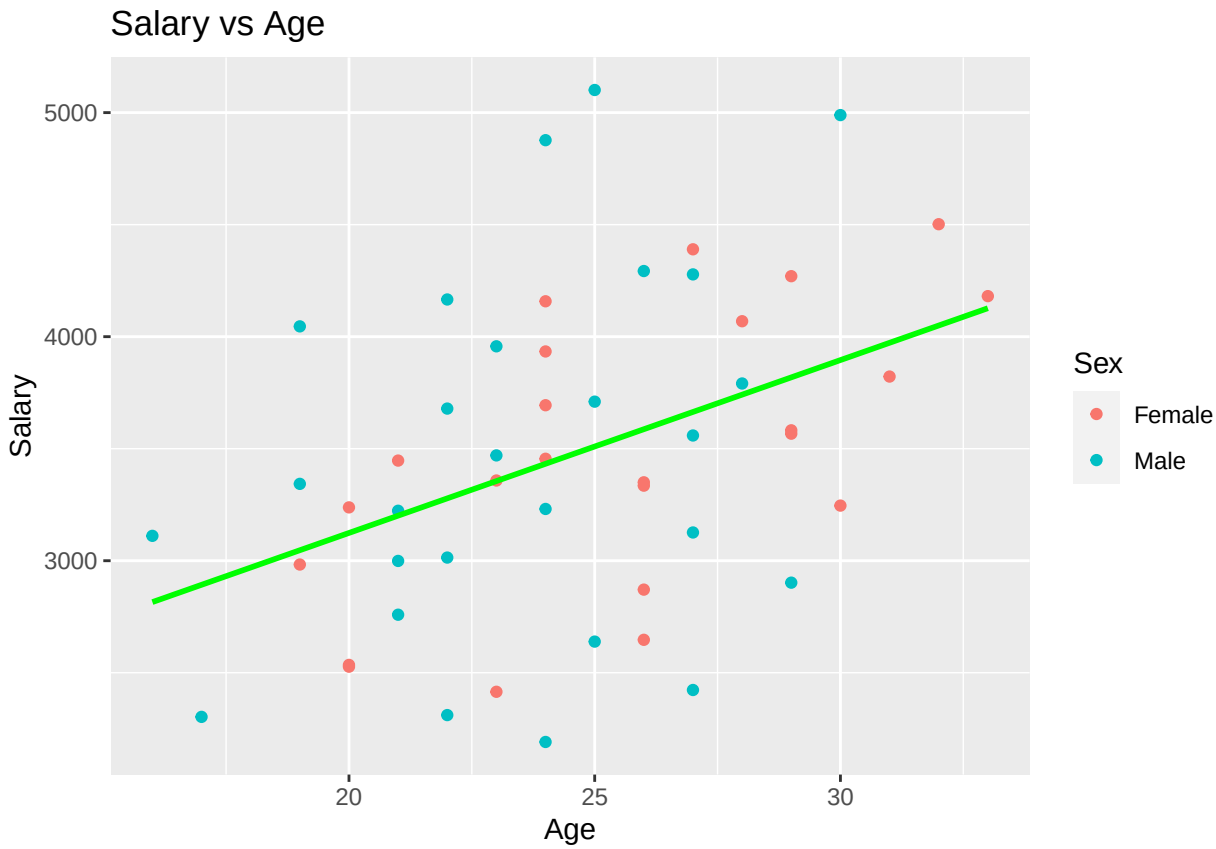
```
## [1] 0.9965094
```

Expense and salary appear to be very strongly correlated, as evidenced by the correlation coefficient and the near perfect linear relationship seen in the scatterplot

Salary vs Age

```
ggplot(df, aes(x=Age, y=Salary)) + geom_point(aes(color = factor(Sex))) + stat_smooth(method = "lm", co
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
boxplot(Salary~Age,data=df, main="Salary Data by Age", xlab="Age", ylab="Salary", col=4)
```

```
cor(df$Salary, df$Age)
```

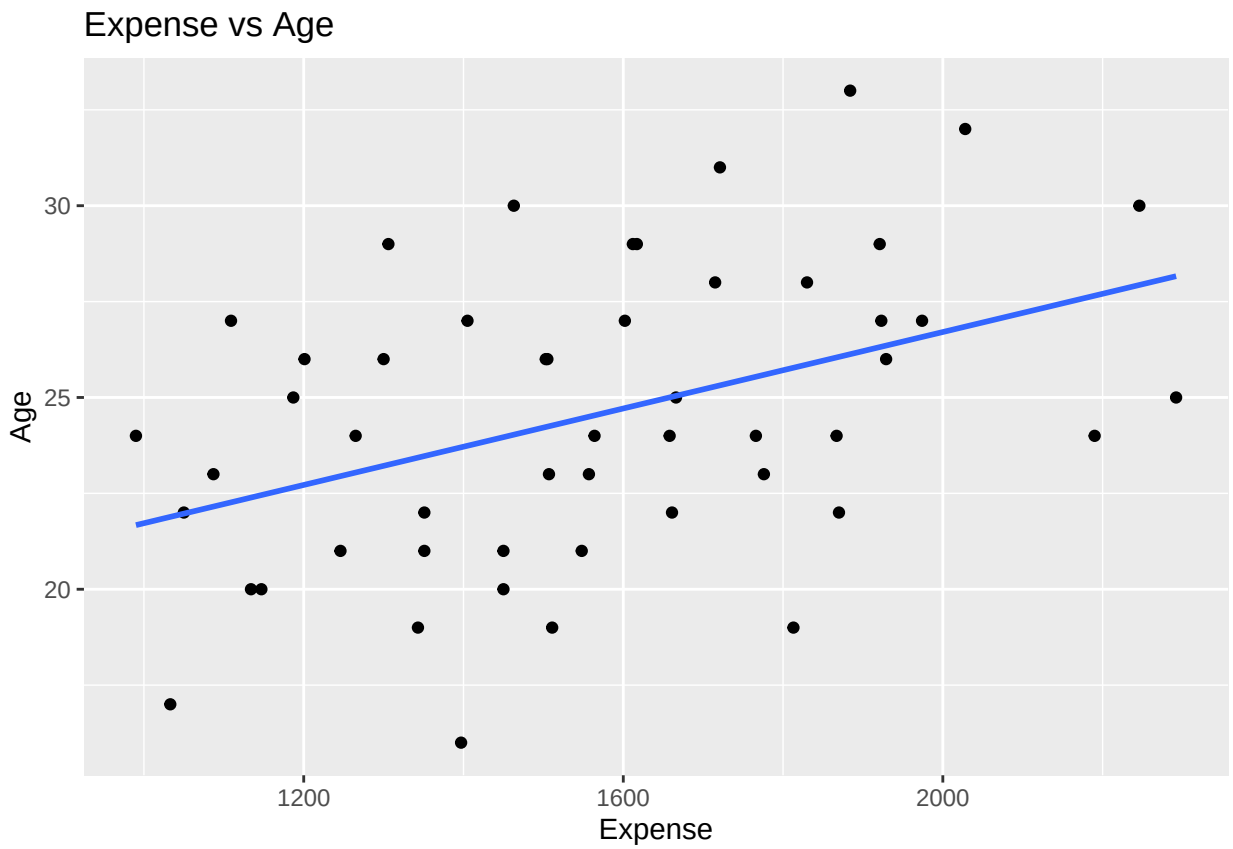
```
## [1] 0.4136556
```

Salary and age are weakly, positively correlated.

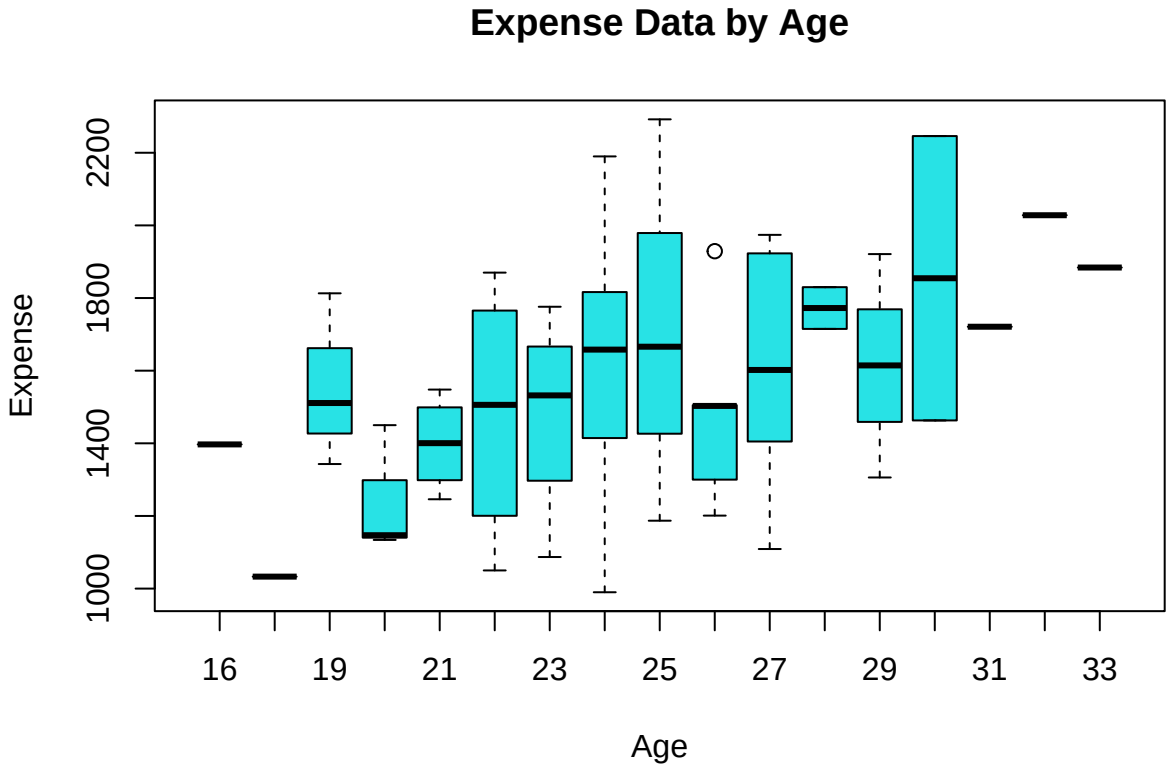
Expense vs Age

```
ggplot(df, aes(x=Expense, y=Age)) + geom_point() + stat_smooth(method = "lm", se= FALSE, size = 1) + lab
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
boxplot(Expense~Age,data=df, main="Expense Data by Age", xlab="Age", ylab="Expense", col=5)
```



This boxplot looks very similar to the boxplot of salary by age, perhaps due to correlation between those variables.

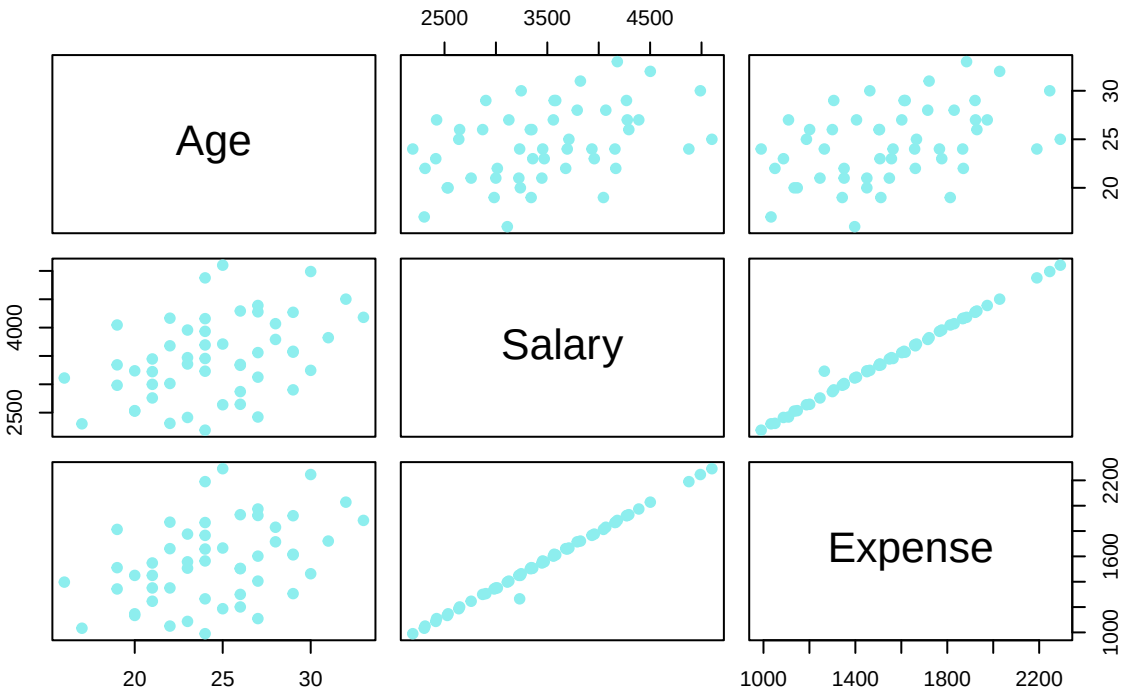
```
cor(df$Expense, df$Age)
```

```
## [1] 0.4189089
```

Expense and age have a somewhat weak, positive correlation, similar to salary and age

Scatterplot Matrix of Age, Salary, and Expense

```
plot(df[3:5], col="darkslategray2", main = "Scatterplot Matrix of Data", pch=19)
```



It's clear from the scatterplot matrix that expense and salary are very closely correlated, while age is not very strongly correlated with either salary or expense. All the relationships are positive