Predictive Analytics

Jatinder Singh Malhi

06/06/2021

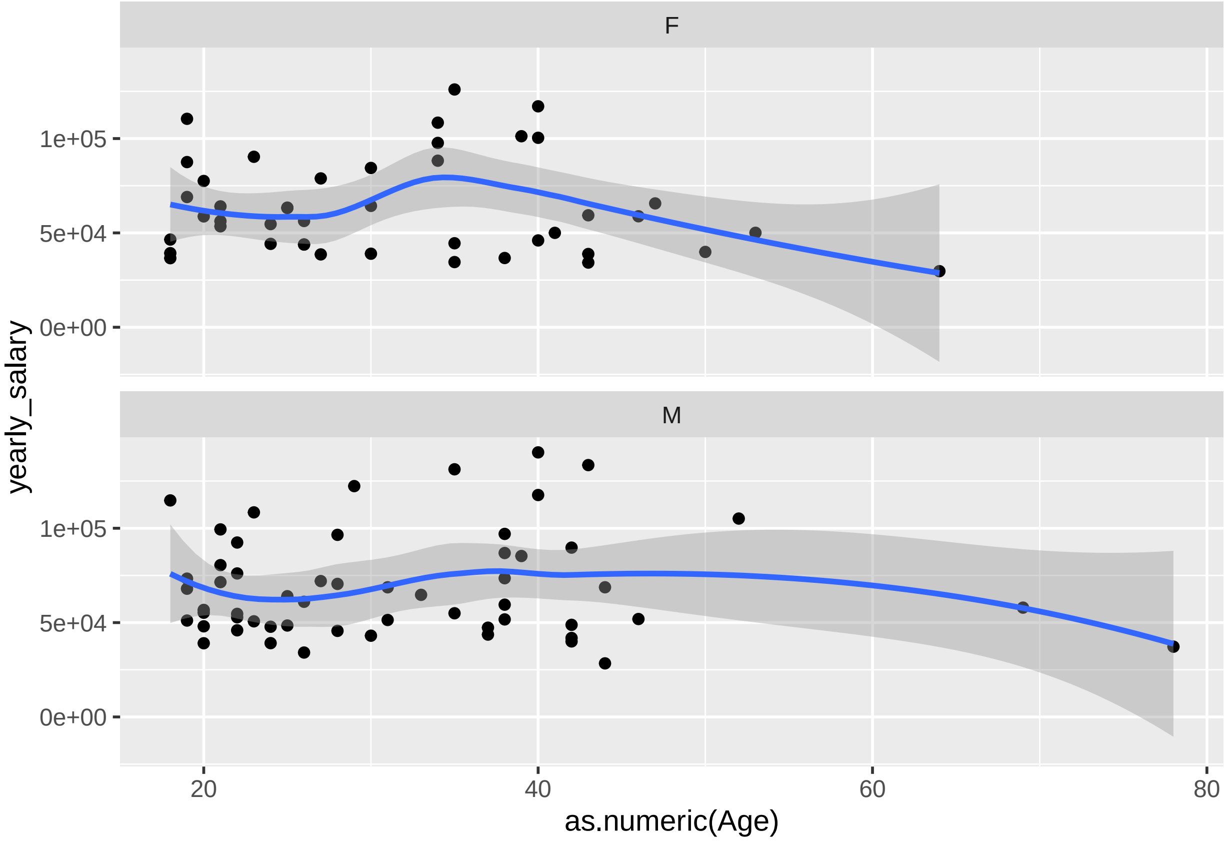
```r
anz_data <- read.csv("ANZ_data_set.csv",comment.char = "#")
head(anz_data)
```

```
##       status card_present_flag bpay_biller_code        account currency
## 1 authorized                 1                  ACC-1598451071      AUD
## 2 authorized                 0                  ACC-1598451071      AUD
## 3 authorized                 1                  ACC-1222300524      AUD
## 4 authorized                 1                  ACC-1037050564      AUD
## 5 authorized                 1                  ACC-1598451071      AUD
## 6     posted                NA                  ACC-1608363396      AUD
##        long_lat txn_description                          merchant_id
## 1 153.41 -27.95             POS 81c48296-73be-44a7-befa-d053f48ce7cd
## 2 153.41 -27.95       SALES-POS 830a451c-316e-4a6a-bf25-e37caedca49e
## 3 151.23 -33.94             POS 835c231d-8cdf-4e96-859d-e9d571760cf0
## 4 153.10 -27.66       SALES-POS 48514682-c78a-4a88-b0da-2d6302e64673
## 5 153.41 -27.95       SALES-POS b4e02c10-0852-4273-b8fd-7b3395e32eb0
## 6 151.22 -33.87         PAYMENT
##   merchant_code first_name balance   date gender age merchant_suburb
## 1            NA      Diana   35.39 1/8/18      F  26         Ashmore
## 2            NA      Diana   21.20 1/8/18      F  26          Sydney
## 3            NA    Michael    5.71 1/8/18      M  38          Sydney
## 4            NA     Rhonda 2117.22 1/8/18      F  40         Buderim
## 5            NA      Diana   17.95 1/8/18      F  26   Mermaid Beach
## 6            NA     Robert 1705.43 1/8/18      M  20
##   merchant_state                   extraction amount
## 1            QLD 2018-08-01T01:01:15.000+0000  16.25
## 2            NSW 2018-08-01T01:13:45.000+0000  14.19
## 3            NSW 2018-08-01T01:26:15.000+0000   6.42
## 4            QLD 2018-08-01T01:38:45.000+0000  40.90
## 5            QLD 2018-08-01T01:51:15.000+0000   3.25
## 6                2018-08-01T02:00:00.000+0000 163.00
##                      transaction_id   country    customer_id merchant_long_lat
## 1 a623070bfead4541a6b0fff8a09e706c Australia CUS-2487424745     153.38 -27.99
## 2 13270a2a902145da9db4c951e04b51b9 Australia CUS-2487424745     151.21 -33.87
## 3 feb79e7ecd7048a5a36ec889d1a94270 Australia CUS-2142601169     151.21 -33.87
## 4 2698170da3704fd981b15e64a006079e Australia CUS-1614226872     153.05 -26.68
## 5 329adf79878c4cf0aeb4188b4691c266 Australia CUS-2487424745     153.44 -28.06
## 6 1005b48a6eda4ffd85e9b649dc9467d3 Australia CUS-2688605418
##   movement
## 1    debit
## 2    debit
## 3    debit
## 4    debit
## 5    debit
## 6    debit
```

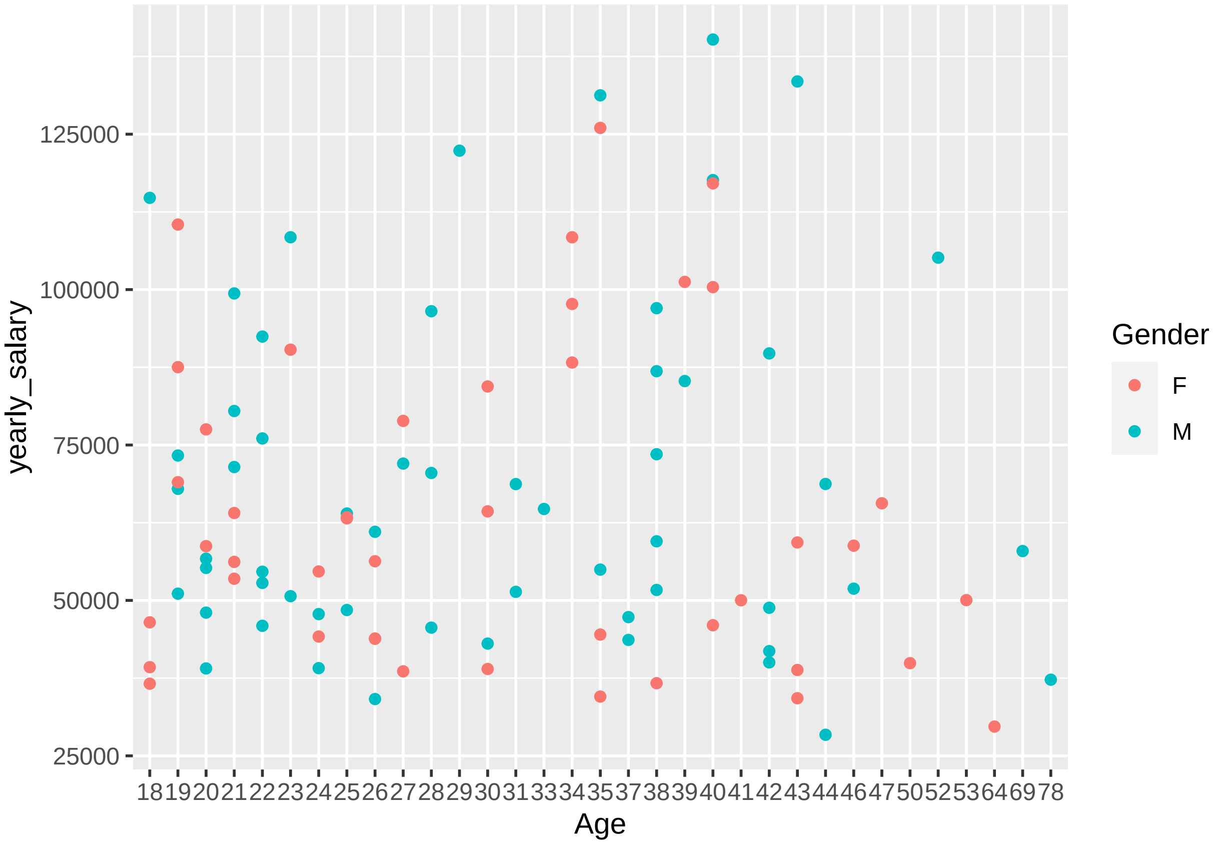
```
unique_id <- unique(anz_data$customer_id)
uniqu_txn <- anz_data[anz_data$txn_description == "PAY/SALARY",]
emptyyy_df <- data.frame()
for (id in unique_id) {
  totalsalary <- 0
  freq <- 0
  for (rw in 1:nrow(uniqu_txn)) {
    if (id == uniqu_txn[rw,"customer_id"])
    {
      totalsalary = totalsalary+uniqu_txn[rw,"amount"]
      freq <- freq +1
      name <- uniqu_txn[rw, "first_name"]
      age <- uniqu_txn[rw,"age"]
      gender <- uniqu_txn[rw,"gender"]
      merchant <- uniqu_txn[rw, "merchant_state"]
    }
  }
  vect <- c(id, name,age, totalsalary,gender, freq, merchant)
  emptyyy_df <-  rbind(emptyyy_df,vect)
}
colnames(emptyyy_df) <- c("ID", "Name", "Age", "Three_Months_Salary", "Gender","Frequency", "State")
emptyyy_df$yearly_salary <- (as.numeric(emptyyy_df$Three_Months_Salary )/92) * 365
```

```
ggplot(data = emptyyy_df,mapping = aes(x= as.numeric(Age), y = yearly_salary),color = "blue")+geom_poin
```

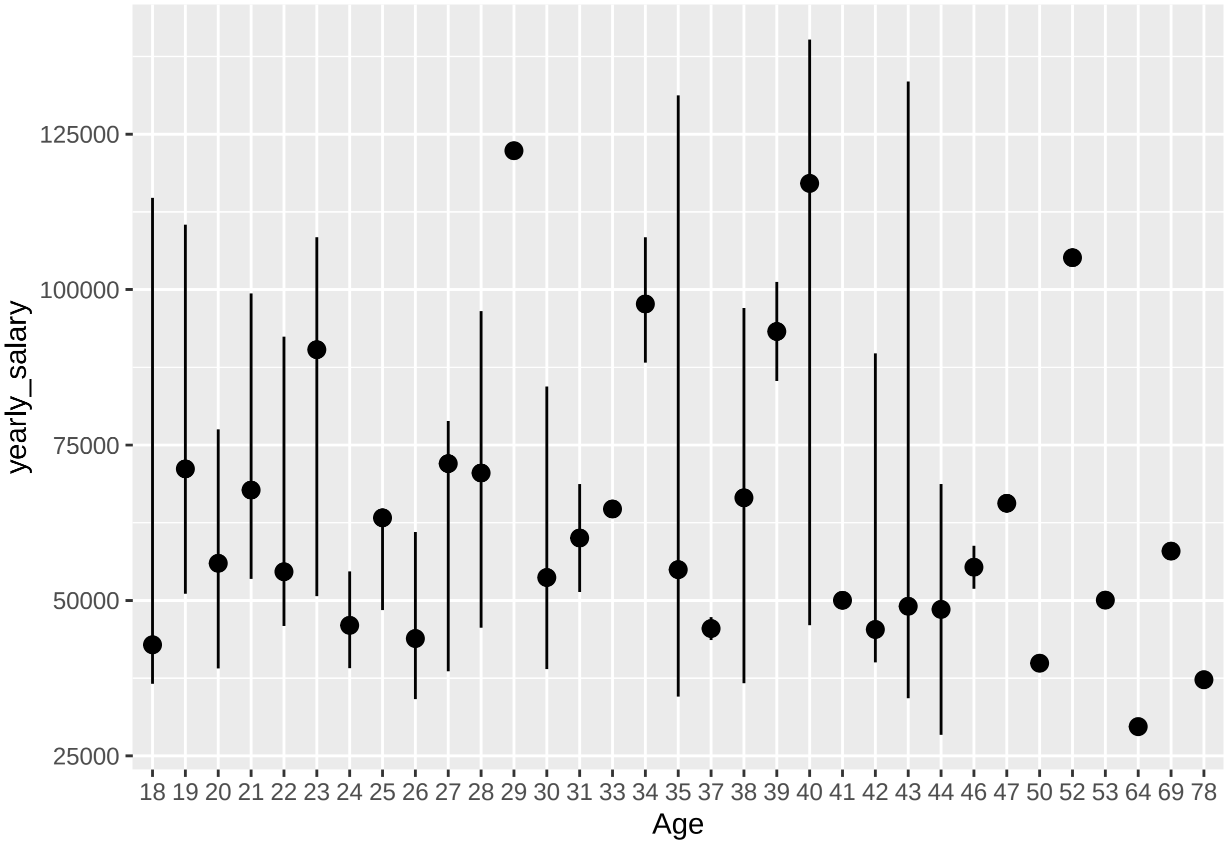
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



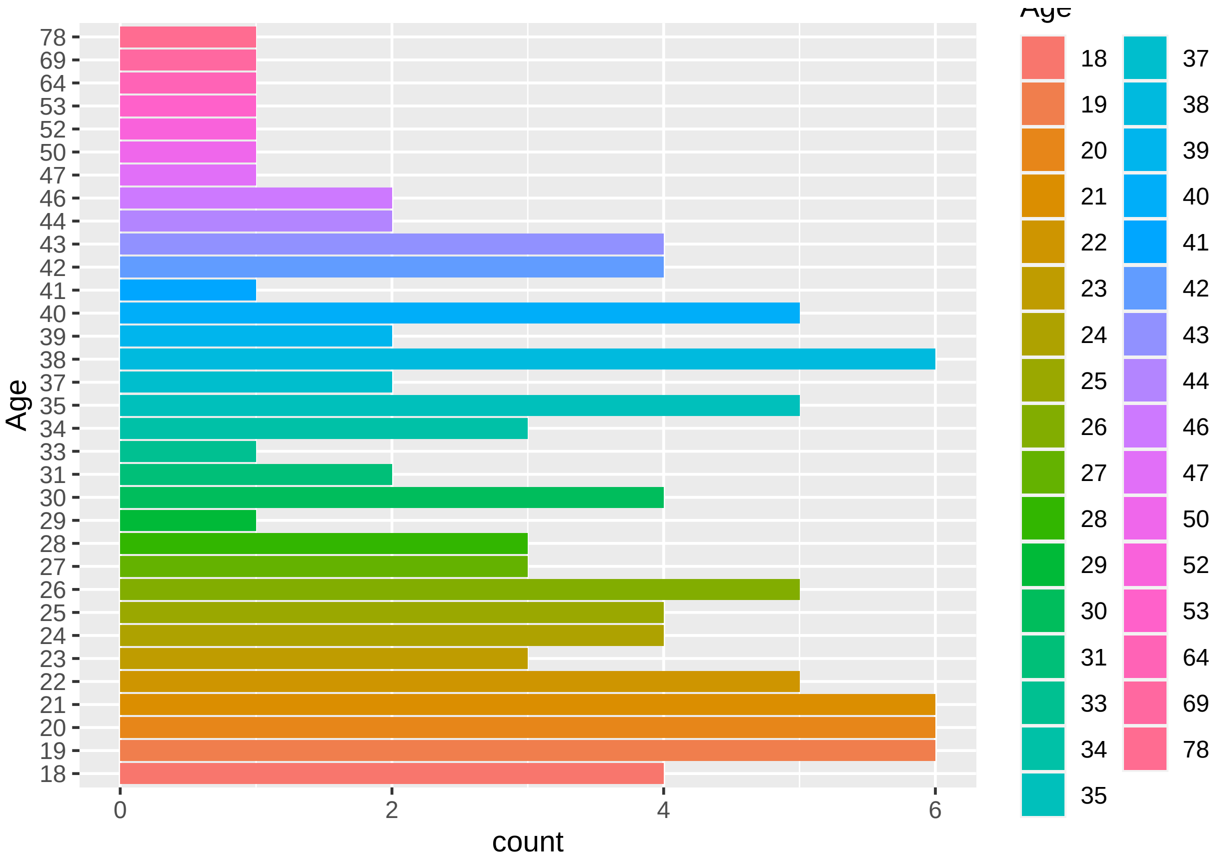
```
ggplot(data = emptyyy_df,mapping = aes(x=Age,y = yearly_salary))+geom_point(mapping = aes(color = Gende
```



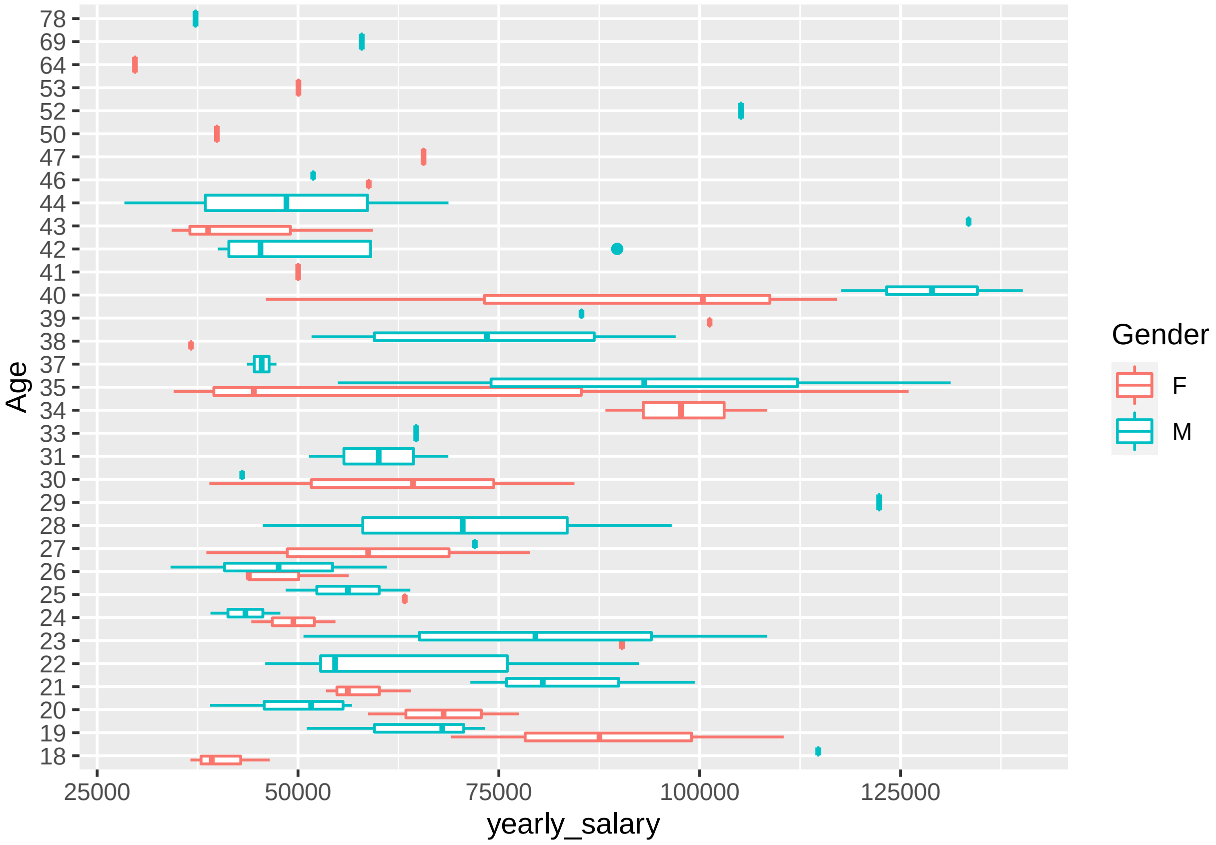
```
ggplot(data =  emptyyy_df)+ geom_pointrange(mapping = aes(x = Age, y = yearly_salary), stat = "summary"
```



```
ggplot(data = emptyyy_df)+geom_bar(mapping = aes(x = Age, fill = Age)) + coord_flip()
```



```
ggplot(data = emptyyy_df, aes(x= Age, y = yearly_salary, color = Gender))+geom_boxplot() + coord_flip()
```



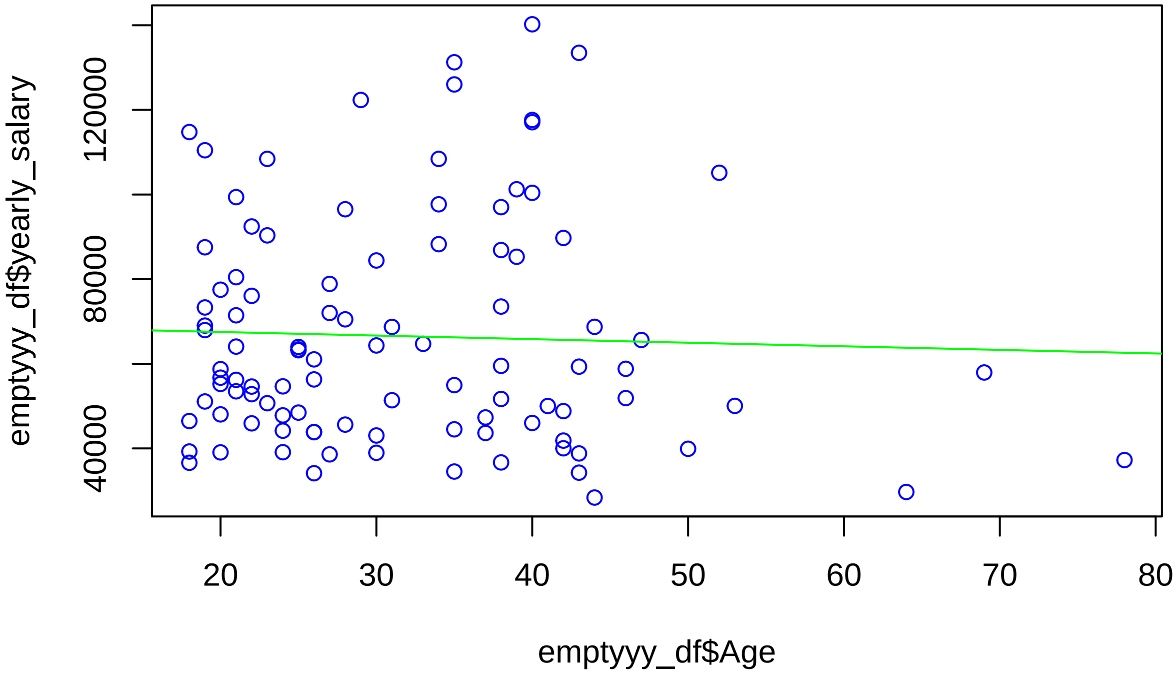
```
# ggplot(data = emptyyy_df, aes(x = as.numeric(Age), y = yearly_salary))+geom_point()+stat_smooth(metho
head(emptyyy_df)
```

```
##                ID    Name Age Three_Months_Salary Gender Frequency State
## 1 CUS-2487424745   Diana  26            14191.38      F        14
## 2 CUS-2142601169 Michael  38            13027.69      M        13
## 3 CUS-1614226872  Rhonda  40            11597.17      F        13
## 4 CUS-2688605418  Robert  20             13921.8      M         6
## 5 CUS-4123612273 Kristin  43            14952.56      F        14
## 6 CUS-3026014945   Tonya  27            19881.05      F         7
##   yearly_salary
## 1      56302.76
## 2      51685.94
## 3      46010.51
## 4      55233.23
## 5      59322.66
## 6      78875.90
```

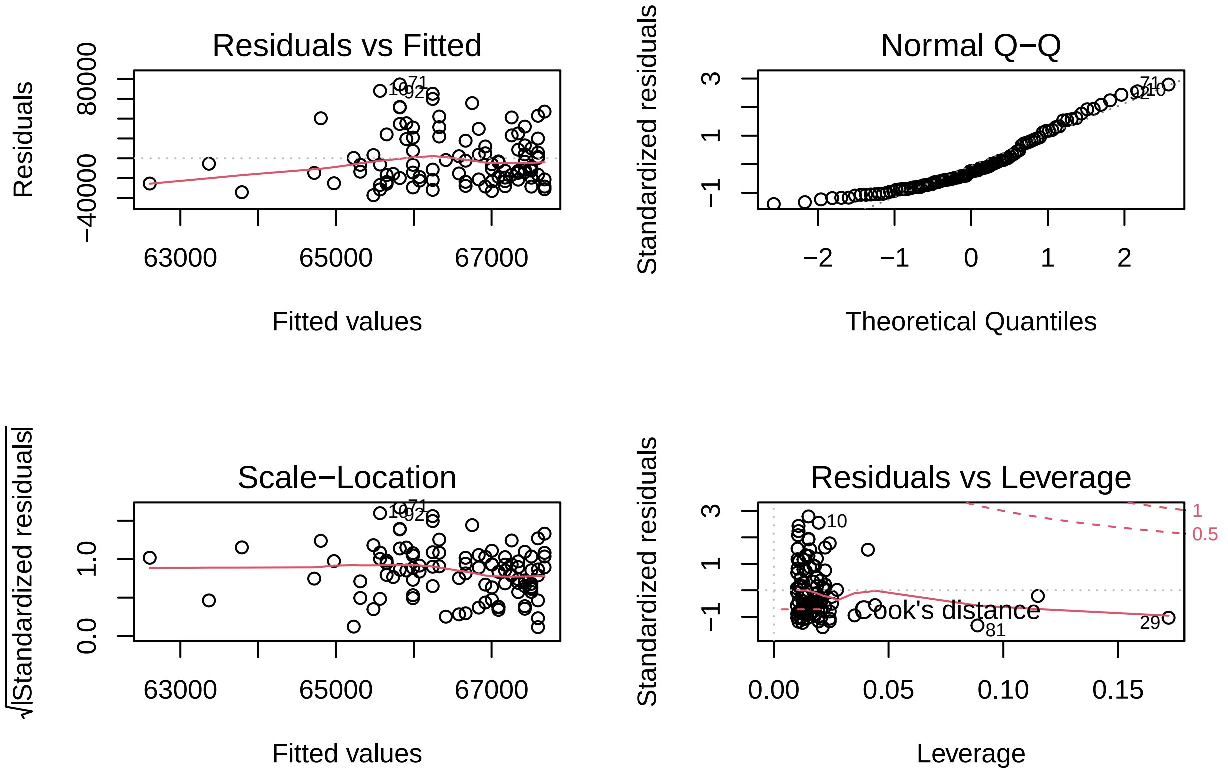
```
linear <- lm(yearly_salary  ~ as.numeric(Age), data = emptyyy_df)
ggpredict(linear)
```

```
## $Age
## # Predicted values of yearly_salary
##
## Age | Predicted |               95% CI
## ----------------------------------------
##  15 |  67934.67 | [58616.98, 77252.37]
##  25 |  67088.93 | [60977.07, 73200.79]
##  30 |  66666.06 | [61338.56, 71993.56]
##  40 |  65820.32 | [59342.92, 72297.71]
##  50 |  64974.58 | [55097.48, 74851.67]
##  55 |  64551.71 | [52672.44, 76430.97]
##  65 |  63705.96 | [47589.00, 79822.92]
##  80 |  62437.35 | [39710.40, 85164.30]
##
## attr(,"class")
## [1] "ggalleffects" "list"
## attr(,"model.name")
## [1] "linear"
```

```
plot(emptyyy_df$Age, emptyyy_df$yearly_salary, col = "blue")
abline(linear, col = "green")
```



```
par(mfrow = (c(2,2)))
plot(linear)
```



```
summary(linear)
```

```
##
## Call:
## lm(formula = yearly_salary ~ as.numeric(Age), data = emptyyy_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37092 -21209  -6492  18159  74403
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     69203.29    7901.15   8.759 5.95e-14 ***
## as.numeric(Age)   -84.57     233.88  -0.362    0.718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26860 on 98 degrees of freedom
## Multiple R-squared:  0.001333,	Adjusted R-squared:  -0.008858
## F-statistic: 0.1308 on 1 and 98 DF,  p-value: 0.7184
```

```
library(modelr)
```

```
##
## Attaching package: 'modelr'
```

```
## The following object is masked from 'package:ggeffects':
##
##     data_grid
```

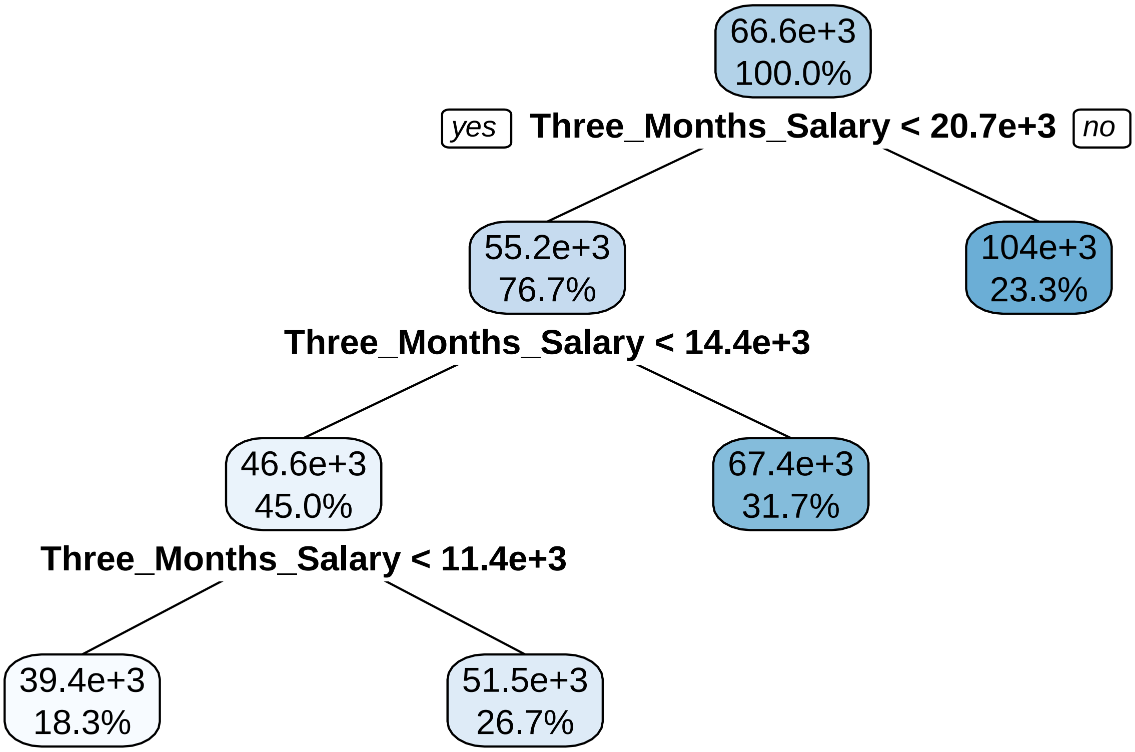
```r
emptyyy_df$Age <- as.numeric(emptyyy_df$Age)
emptyyy_df$Three_Months_Salary <- as.numeric(emptyyy_df$Three_Months_Salary)
emptyyy_df$Gender <- as.numeric(emptyyy_df$Gender)
```

```
## Warning: NAs introduced by coercion
```

```r
cor(emptyyy_df[,c(3,4)])
```

```
##                             Age Three_Months_Salary
## Age                  1.0000000          -0.0365039
## Three_Months_Salary -0.0365039           1.0000000
```

```r
train_Index <- 1:(size = floor(0.60*nrow(emptyyy_df)))
df_train <- emptyyy_df[train_Index,]
df_test <- emptyyy_df[-train_Index,]
Y_Train <- df_train$yearly_salary
Y_Test <- df_test$yearly_salary
df_train <- df_train[,c(3,4)]
df_test <- df_test[,c(3,4)]
XY_train <- cbind(as.data.frame(df_train),Y = Y_Train)
XY_test <- cbind(as.data.frame(df_test), Y = Y_Test)
t1 <- rpart(Y ~., data= XY_train, method = "anova")
rpart.plot(t1, tweak = 1.1, fallen.leaves = FALSE, digits = 3)
```



```r
Y_predi <- predict(t1, XY_test, method = "anova")
R2 <- rsquare(t1, data = XY_train)
RMSE <- rmse(t1, data = XY_train)
MAE <- mae(t1, data = XY_train)
R22 <- rsquare(t1, data = XY_test)
RMSE2 <- rmse(t1, data = XY_test)
MAE2 <-mae(t1, data = XY_test)
```

```r
train_test_comparison <- (cbind(R2, RMSE, MAE))
train_test_comparison1 <- (cbind(R22, RMSE2, MAE2))
tmt <- as.data.frame(rbind(train_test_comparison,train_test_comparison1))
row.names(tmt) <- c("Train_model", "Test_model")
knitr::kable(tmt)
```

	R2	RMSE	MAE
Train_model	0.9064550	7342.722	5428.120
Test_model	0.8575961	11358.818	7889.052