Univariate Analysis For Banking Data

Ada Lazuli

2022-09-05

Contents

Libraries

```
library(e1071)
library(ggplot2)
library(dplyr)
```

Helper Functions

```
base_hist <- function(df, x, target_name){
  title <- paste("Histogram of ", target_name)
  ggplot(df, aes(x = x)) + geom_histogram(alpha = 0.4) +
    ggtitle(title) +
    xlab(target_name) +
    ylab("Frequency Count")
}

zero_plot <- function(df, x, target_name) {
  title <- paste("Zero Plot of ", target_name)
  ggplot(df, aes(x = x, y = 0, color =  default.payment.next.month)) +
    geom_point(alpha = 0.45) +
    ggtitle(title) +
    xlab(target_name) +
    theme(axis.ticks.y = element_blank(), axis.text.y = element_blank(),
          axis.title.y = element_blank()) +
```

```
    labs(color = "Defaults")
}

boxplot_comp <- function(df, x, target_name) {
  title <- paste("Boxplot of ", target_name)
  ggplot(df, aes(x = x, fill = default.payment.next.month)) +
    geom_boxplot(alpha = 0.45) +
    ggtitle(title) +
    xlab(target_name) +
    theme(axis.ticks.y = element_blank(), axis.text.y = element_blank(),
          axis.title.y = element_blank()) +
    labs(fill = "Defaults")
}

bar_comp_plot <- function(df, x, target_name) {
  title <- paste("Barplot of ", target_name)
  ggplot(df, aes(x = x, fill =  default.payment.next.month, color = default.payment.next.month)) +
    geom_bar(alpha = 0.5, position = "dodge") +
    ggtitle(title) +
    xlab(target_name) +
    ylab("Frequency Count") +
    labs(fill = "Defaults", color = "Defaults")
}
```

Data Loading

```
df <- read.csv("credit card default data set.csv")
df$default.payment.next.month <- factor(df$default.payment.next.month, labels = c("No", "Yes"))
```

Univariate Analysis

Overview

The dataset has **30000** rows and **25** features. The names of the features are:

ID, LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6, default.payment.next.month

```
glimpse(df)
```

```
## Rows: 30,000
## Columns: 25
## $ ID                        <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ~
## $ LIMIT_BAL                 <int> 20000, 120000, 90000, 50000, 50000, 50000, ~
## $ SEX                       <int> 2, 2, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2, 2, 1, 1~
## $ EDUCATION                 <int> 2, 2, 2, 2, 2, 1, 1, 2, 3, 3, 3, 1, 2, 2, 1~
## $ MARRIAGE                  <int> 1, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2~
## $ AGE                       <int> 24, 26, 34, 37, 57, 37, 29, 23, 28, 35, 34,~
```

```
## $ PAY_0                      <int> 2, -1, 0, 0, -1, 0, 0, 0, 0, -2, 0, -1, -1,~
## $ PAY_2                      <int> 2, 2, 0, 0, 0, 0, 0, -1, 0, -2, 0, -1, 0, 2~
## $ PAY_3                      <int> -1, 0, 0, 0, -1, 0, 0, -1, 2, -2, 2, -1, -1~
## $ PAY_4                      <int> -1, 0, 0, 0, 0, 0, 0, 0, 0, -2, 0, -1, -1, ~
## $ PAY_5                      <int> -2, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, -1, -1, ~
## $ PAY_6                      <int> -2, 2, 0, 0, 0, 0, 0, -1, 0, -1, -1, 2, -1,~
## $ BILL_AMT1                  <int> 3913, 2682, 29239, 46990, 8617, 64400, 3679~
## $ BILL_AMT2                  <int> 3102, 1725, 14027, 48233, 5670, 57069, 4120~
## $ BILL_AMT3                  <int> 689, 2682, 13559, 49291, 35835, 57608, 4450~
## $ BILL_AMT4                  <int> 0, 3272, 14331, 28314, 20940, 19394, 542653~
## $ BILL_AMT5                  <int> 0, 3455, 14948, 28959, 19146, 19619, 483003~
## $ BILL_AMT6                  <int> 0, 3261, 15549, 29547, 19131, 20024, 473944~
## $ PAY_AMT1                   <int> 0, 0, 1518, 2000, 2000, 2500, 55000, 380, 3~
## $ PAY_AMT2                   <int> 689, 1000, 1500, 2019, 36681, 1815, 40000, ~
## $ PAY_AMT3                   <int> 0, 1000, 1000, 1200, 10000, 657, 38000, 0, ~
## $ PAY_AMT4                   <int> 0, 1000, 1000, 1100, 9000, 1000, 20239, 581~
## $ PAY_AMT5                   <int> 0, 0, 1000, 1069, 689, 1000, 13750, 1687, 1~
## $ PAY_AMT6                   <int> 0, 2000, 5000, 1000, 679, 800, 13770, 1542,~
## $ default.payment.next.month <fct> Yes, Yes, No, No, No, No, No, No, No, No, N~
```

ID

The ID feature appears to be an index with a minimum of **1** and maximum of **30000**.

LIMIT_BAL (X1)

X1 is the first feature for predictive modeling and consists for the amount of credit given that accounts for both the individual credit and family credit.

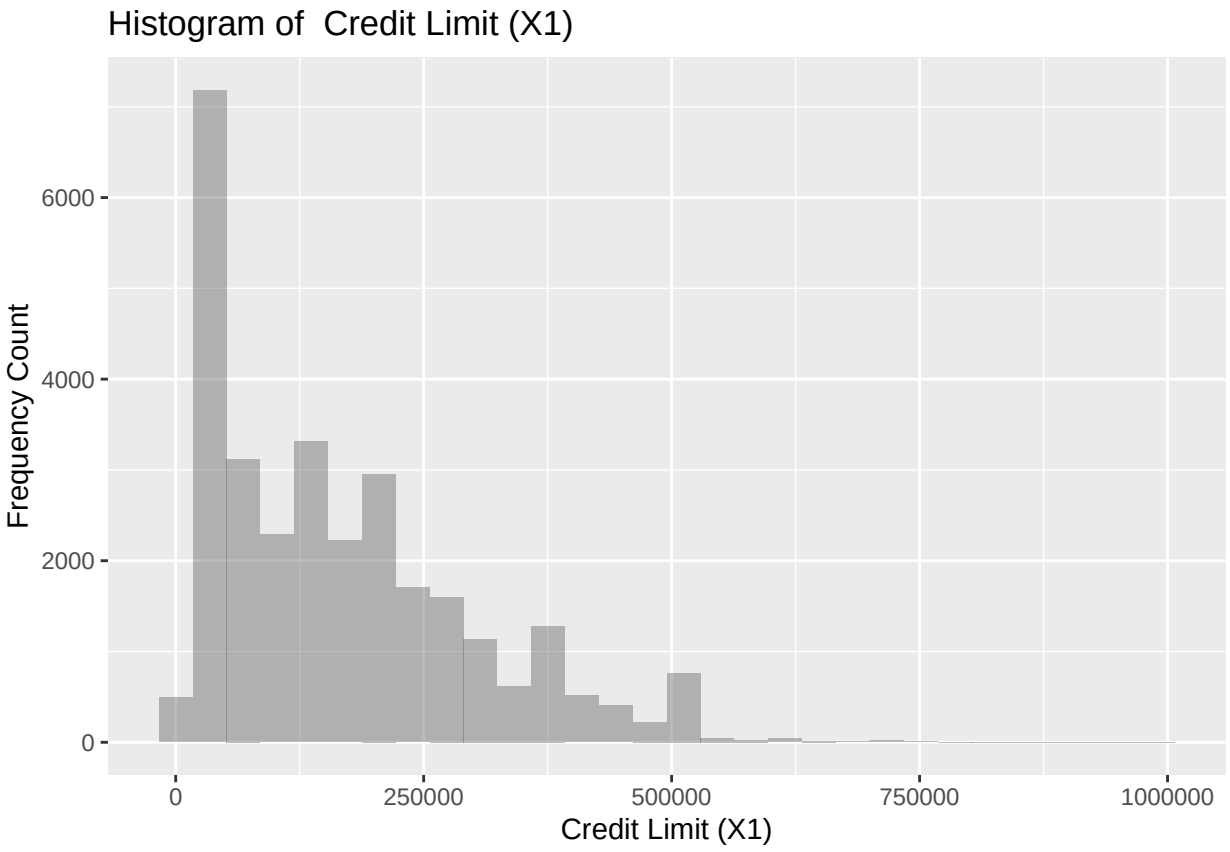
The variable appears to consist of integers, with a maximum of **1000000** and a minimum of **10000**.

```
summary(df$LIMIT_BAL)
```

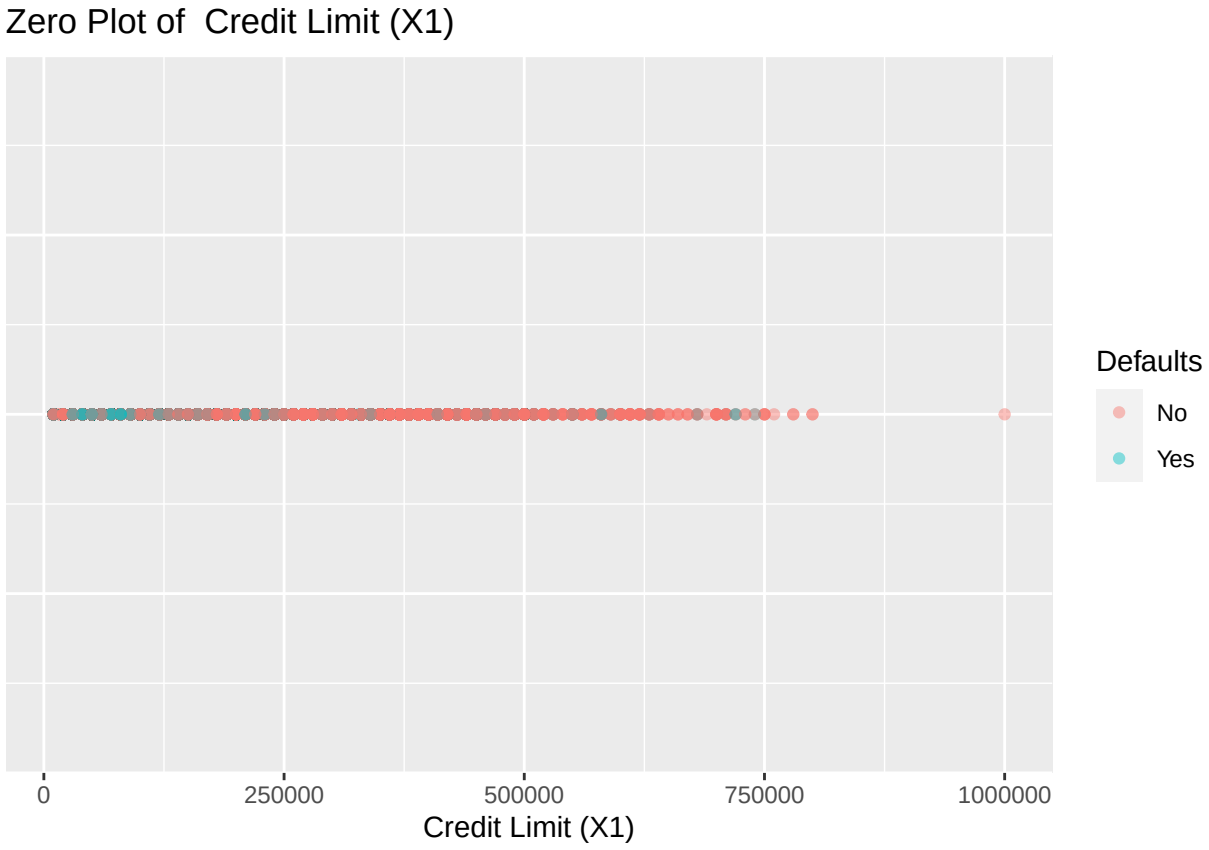
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   50000  140000  167484  240000 1000000
```

Graphical Analysis

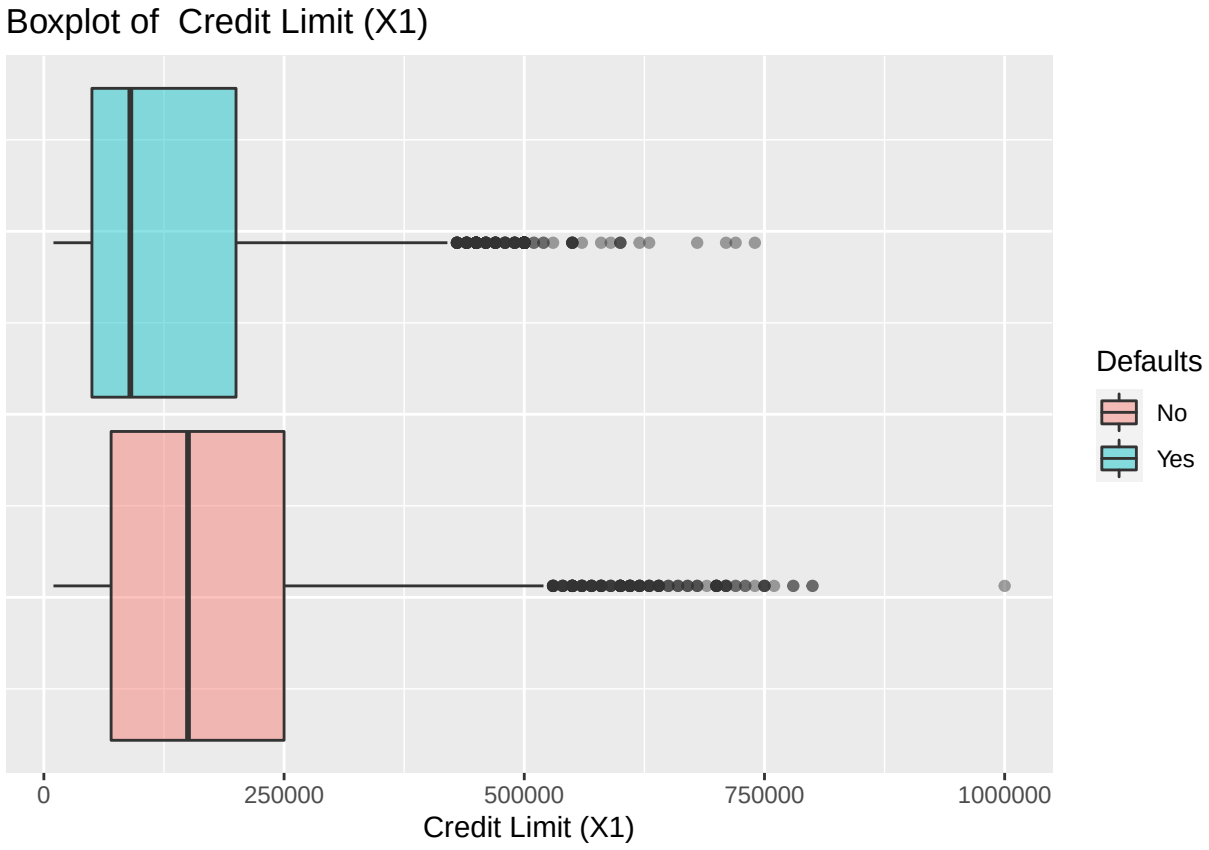
```
x <- df$LIMIT_BAL
name <- "Credit Limit (X1)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



SEX (X2)

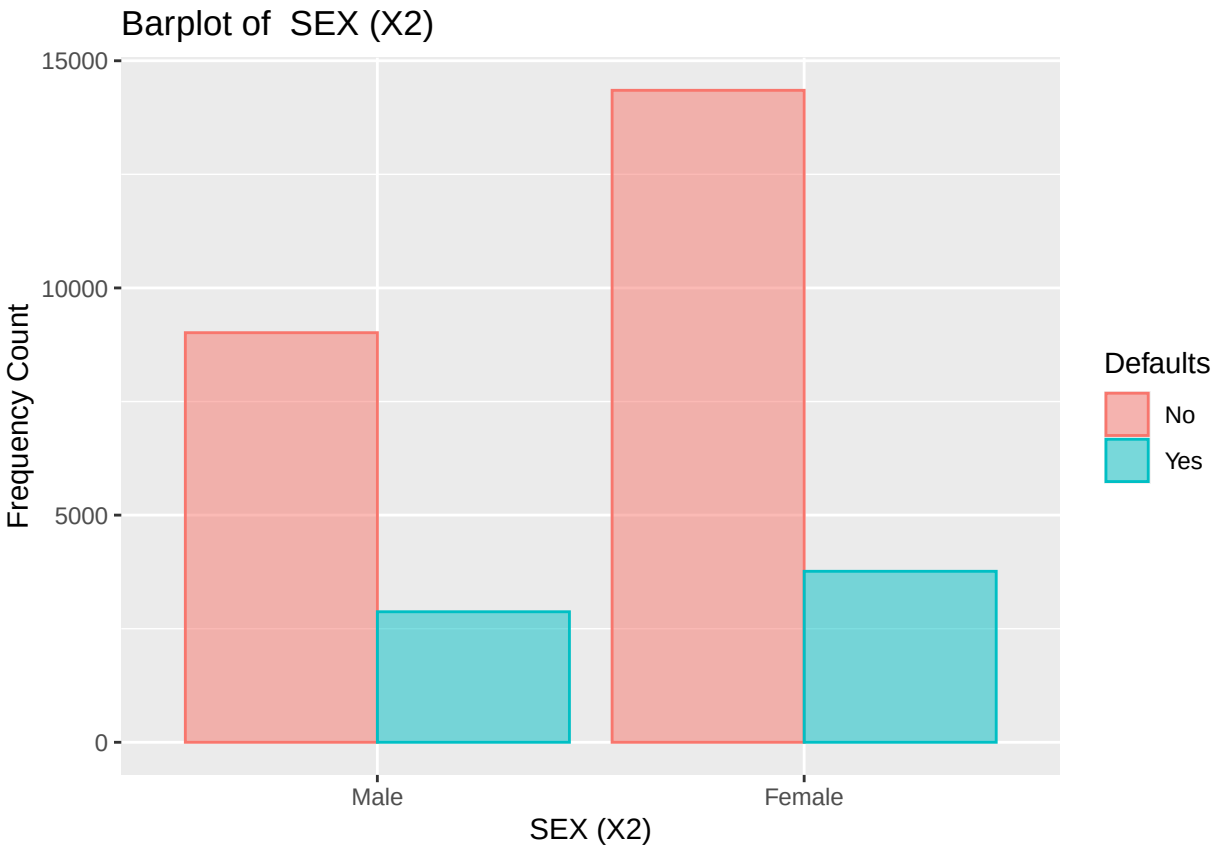
The gender variable refers to the gender of the individual, with a 1 for male and a 2 for female.

```
df$SEX <- factor(df$SEX, labels = c("Male", "Female"))
table(df$SEX)
```

```
##
##   Male Female
##  11888  18112
```

Graphical Analysis

```
x <- df$SEX
name <- "SEX (X2)"
bar_comp_plot(df,x,name)
```



EDUCATION (X3)

The education feature captures the education level of the individual. Possible values are: 1. Graduate School 2. University 3. High School 4. Others

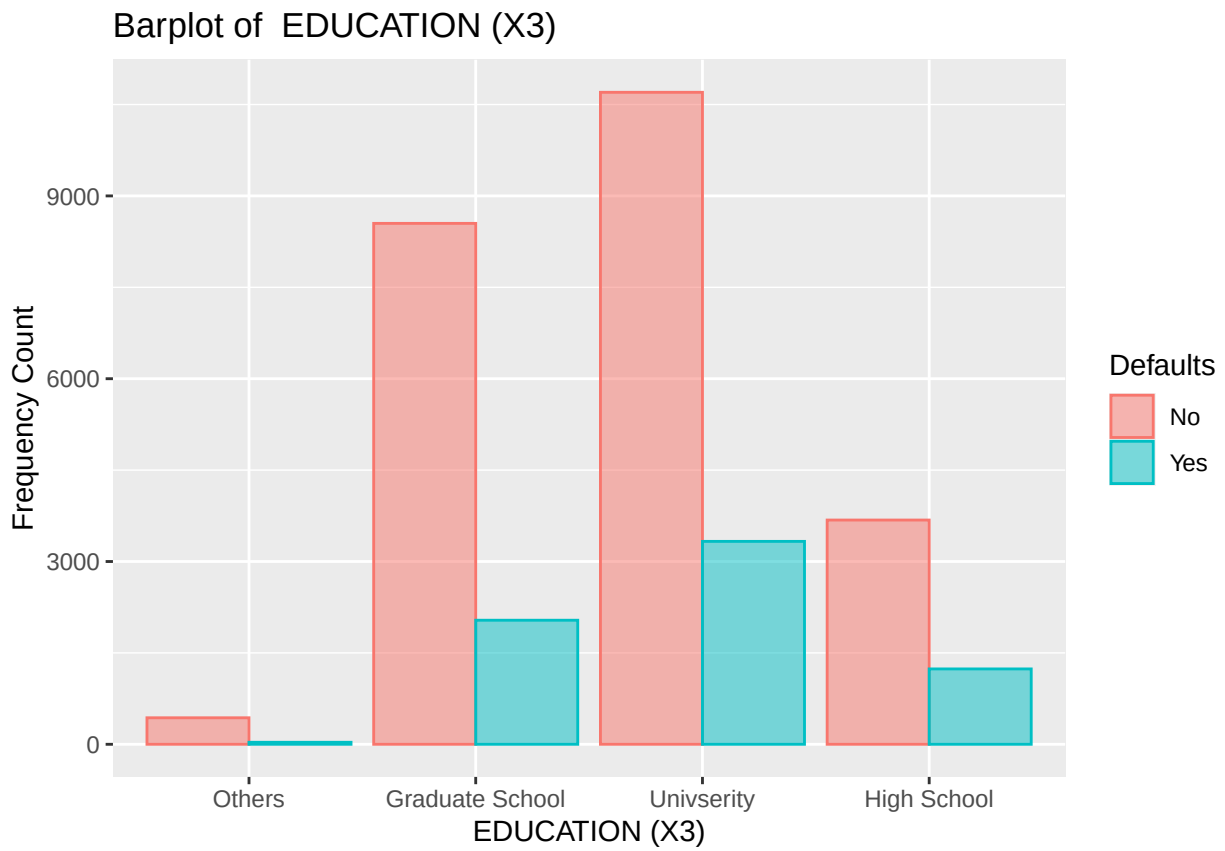
NOTE: Values of 0, 5, and 6 were found in the data and added to the *Others* category.

```
df$EDUCATION <- factor(df$EDUCATION, labels = c("Others", "Graduate School", "Univserity", "High School
table(df$EDUCATION)
```

```
##
##          Others Graduate School      Univserity     High School
##             468           10585           14030            4917
```

Graphical Analysis

```
x <- df$EDUCATION
name <- "EDUCATION (X3)"
bar_comp_plot(df,x,name)
```



MARRIAGE (X4)

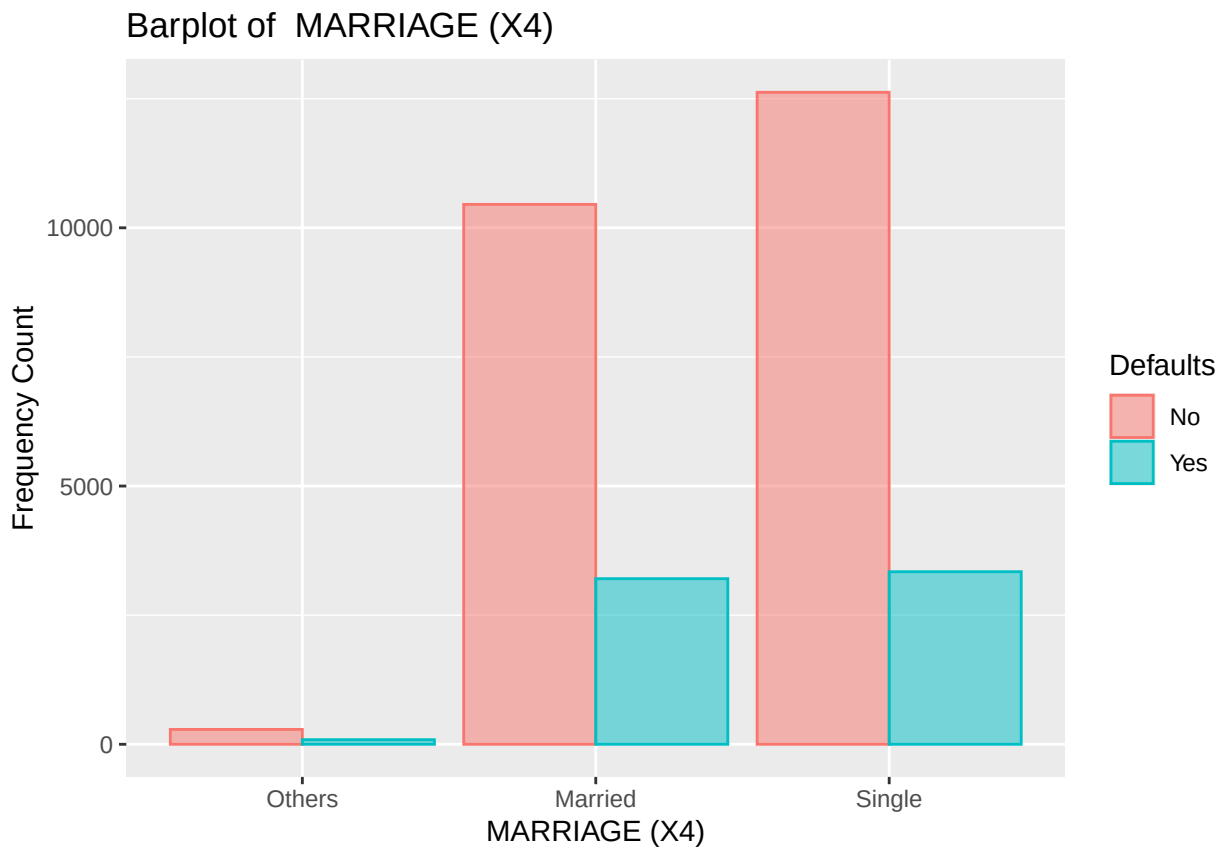
The marriage feature captures the marital status of the individual. Possible values are: 1. Married 2. Single 3. Others

```
df$MARRIAGE <- factor(df$MARRIAGE, labels = c("Others", "Married", "Single","Others"))
table(df$MARRIAGE)
```

```
##
##  Others Married  Single
##     377   13659   15964
```

Graphical Analysis

```
x <- df$MARRIAGE
name <- "MARRIAGE (X4)"
bar_comp_plot(df,x,name)
```



AGE (X5)

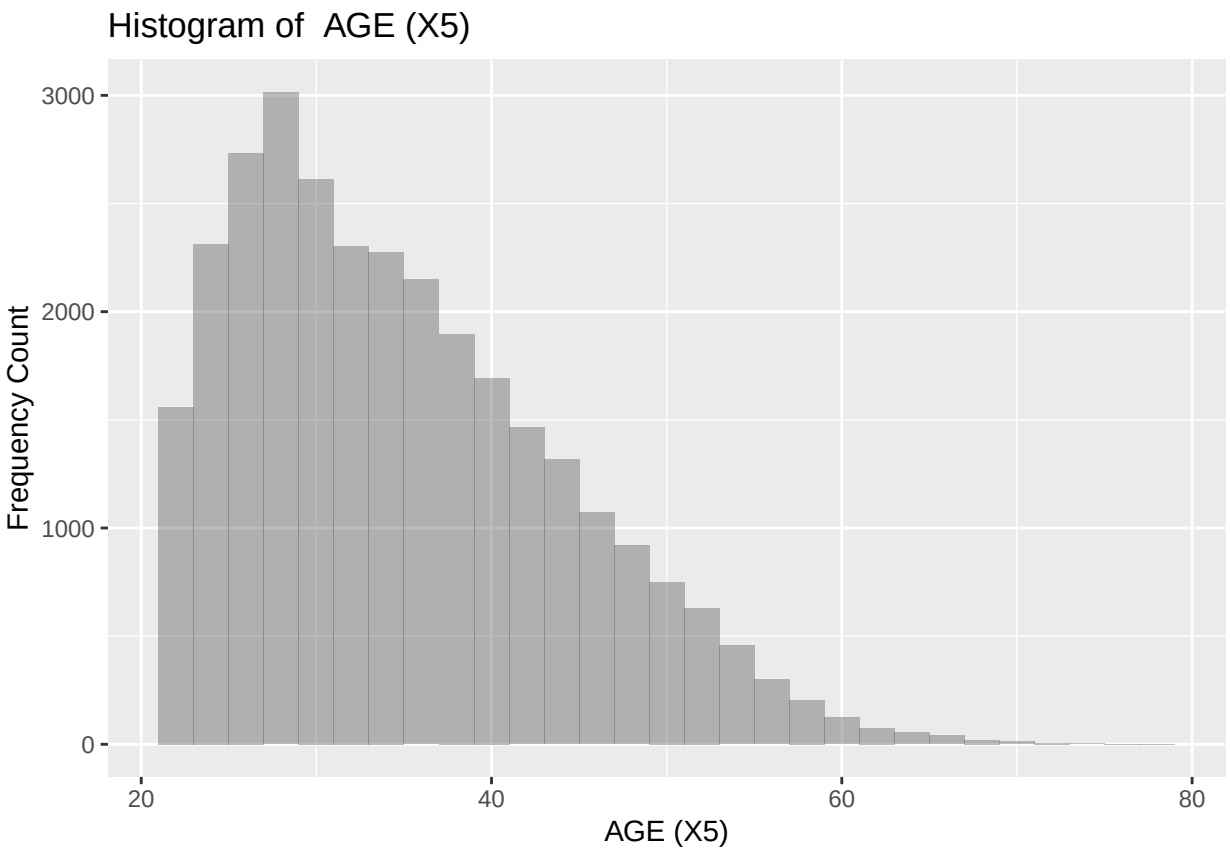
The age feature captures the age of the individual.

```
summary(df$AGE)
```

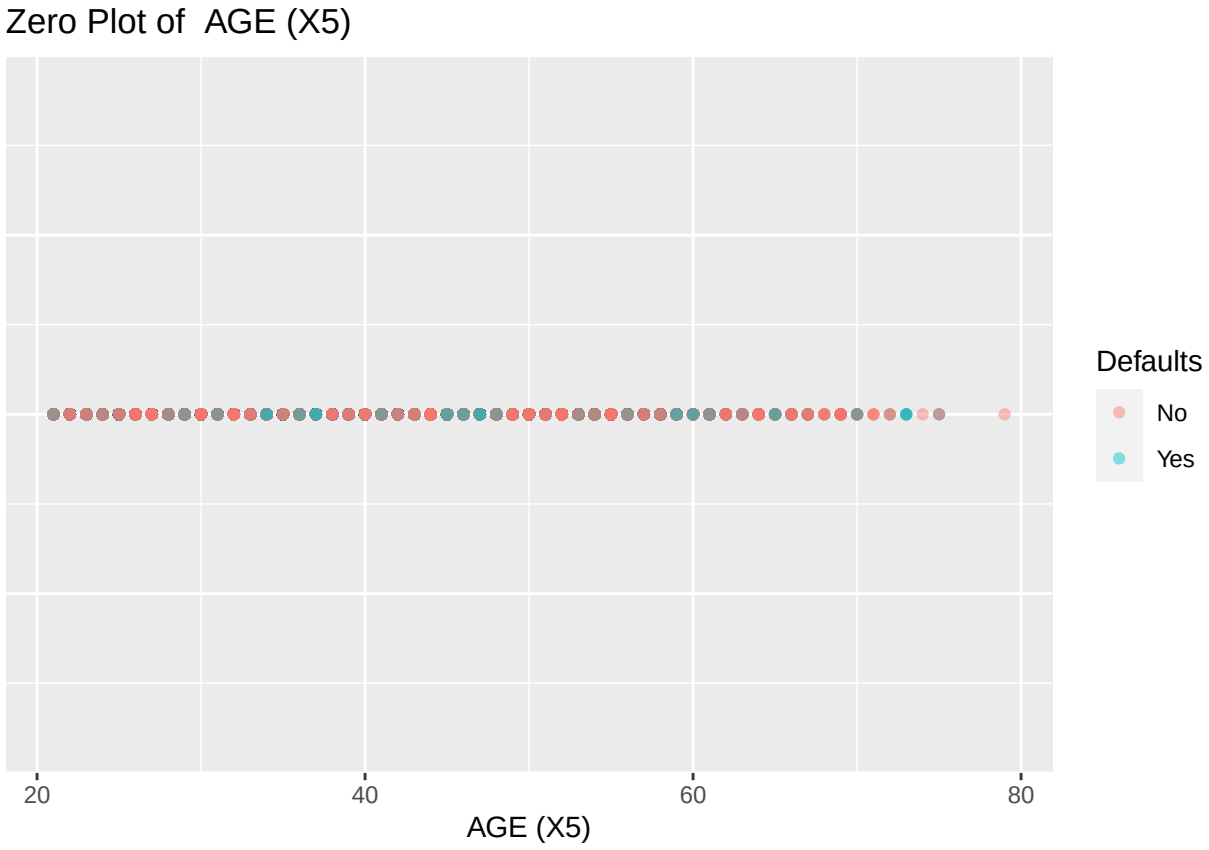
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   28.00   34.00   35.49   41.00   79.00
```

Graphical Analysis

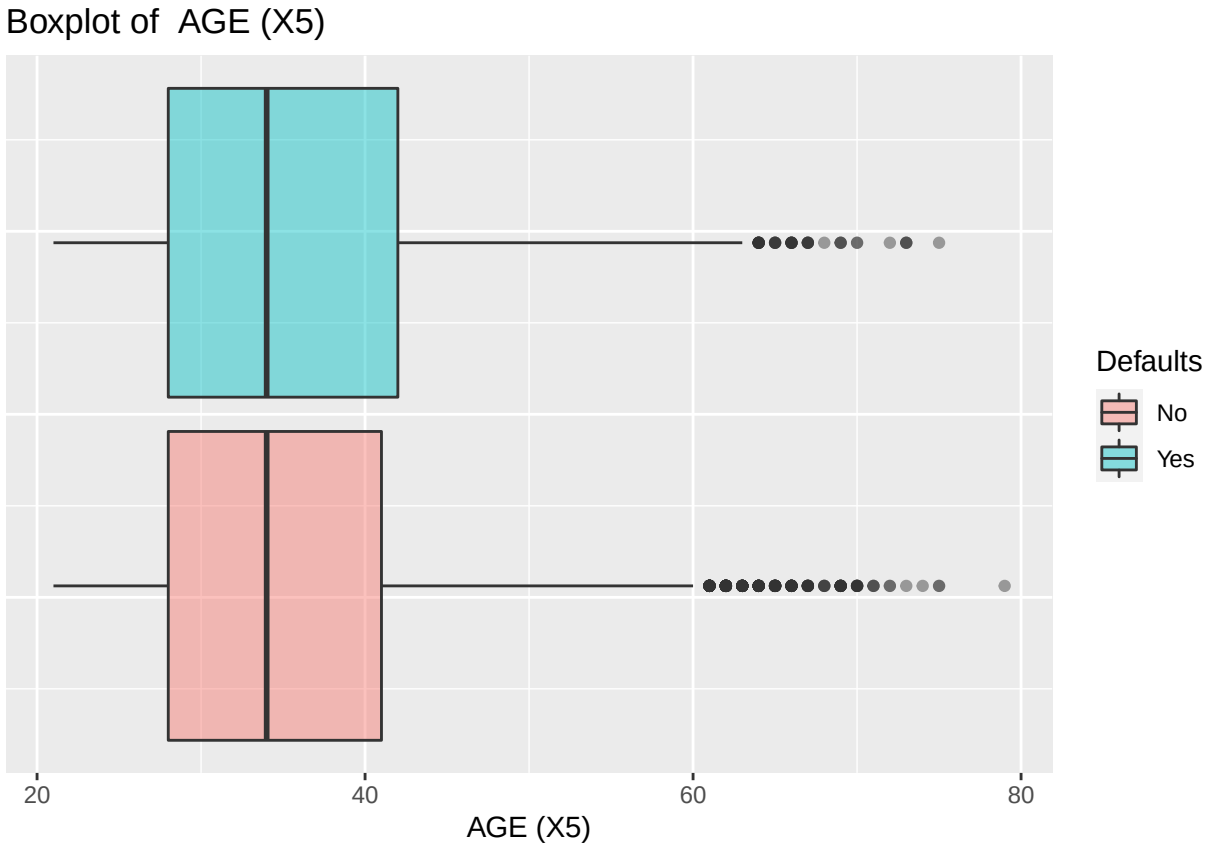
```
x <- df$AGE
name <- "AGE (X5)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_0 (X6)

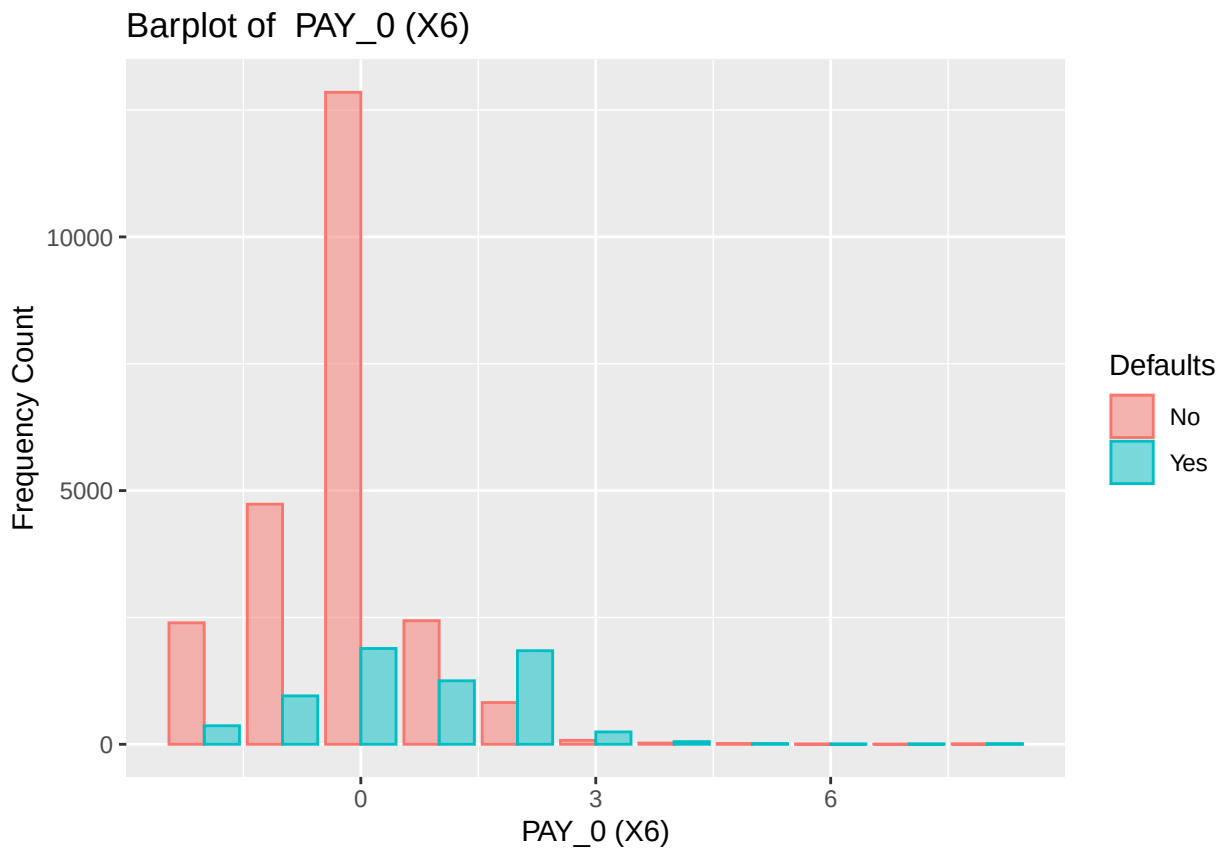
The PAY_0 feature captures the payment status of the individual for the month of **September, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_0)
```

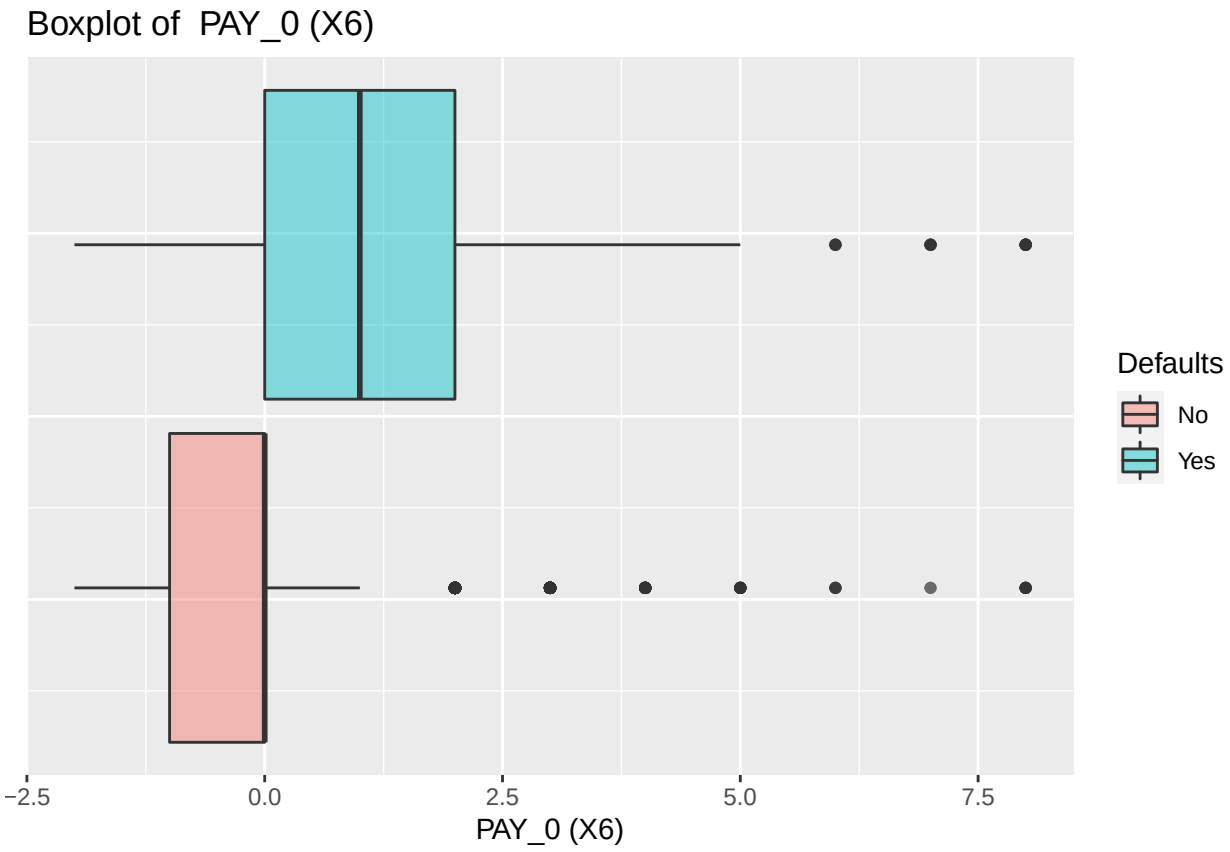
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.0167  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_0
name <- "PAY_0 (X6)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_2 (X7)

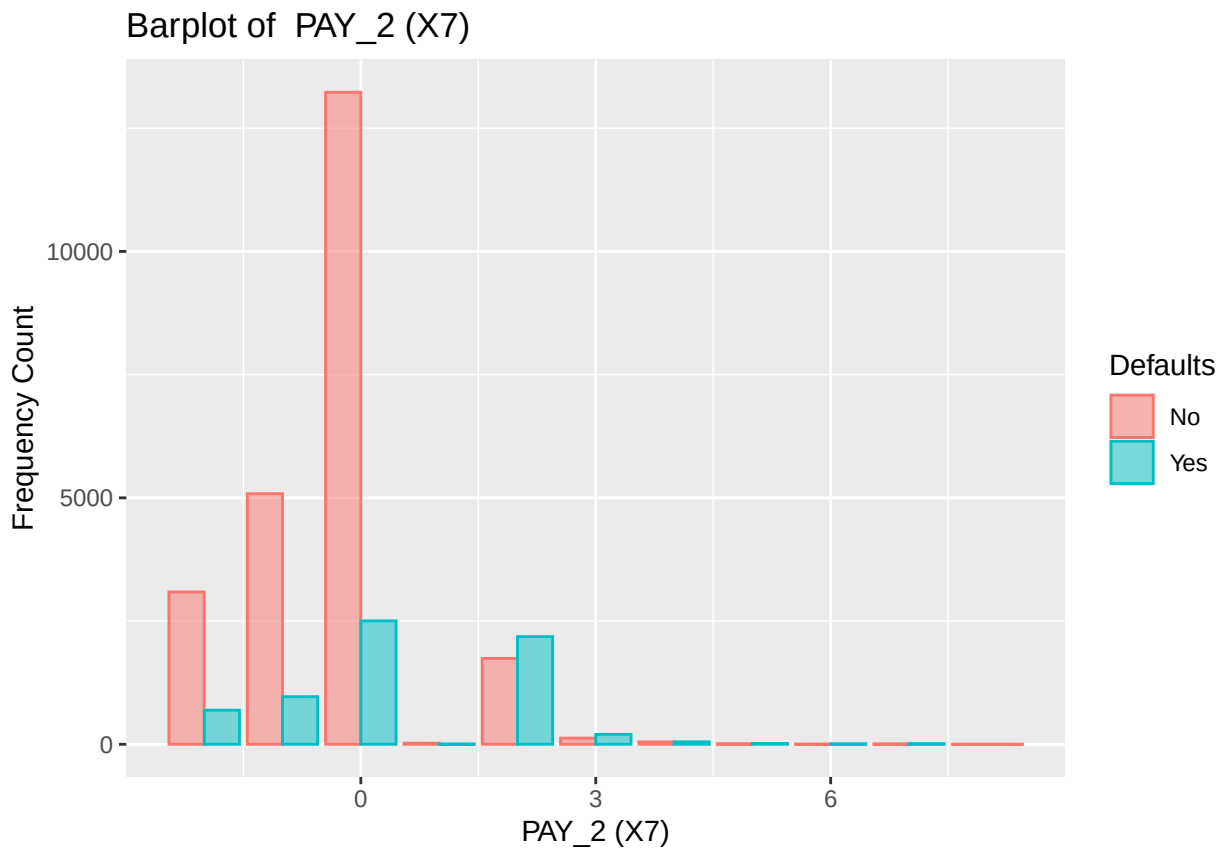
The PAY_1 feature captures the payment status of the individual for the month of **August, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_2)
```

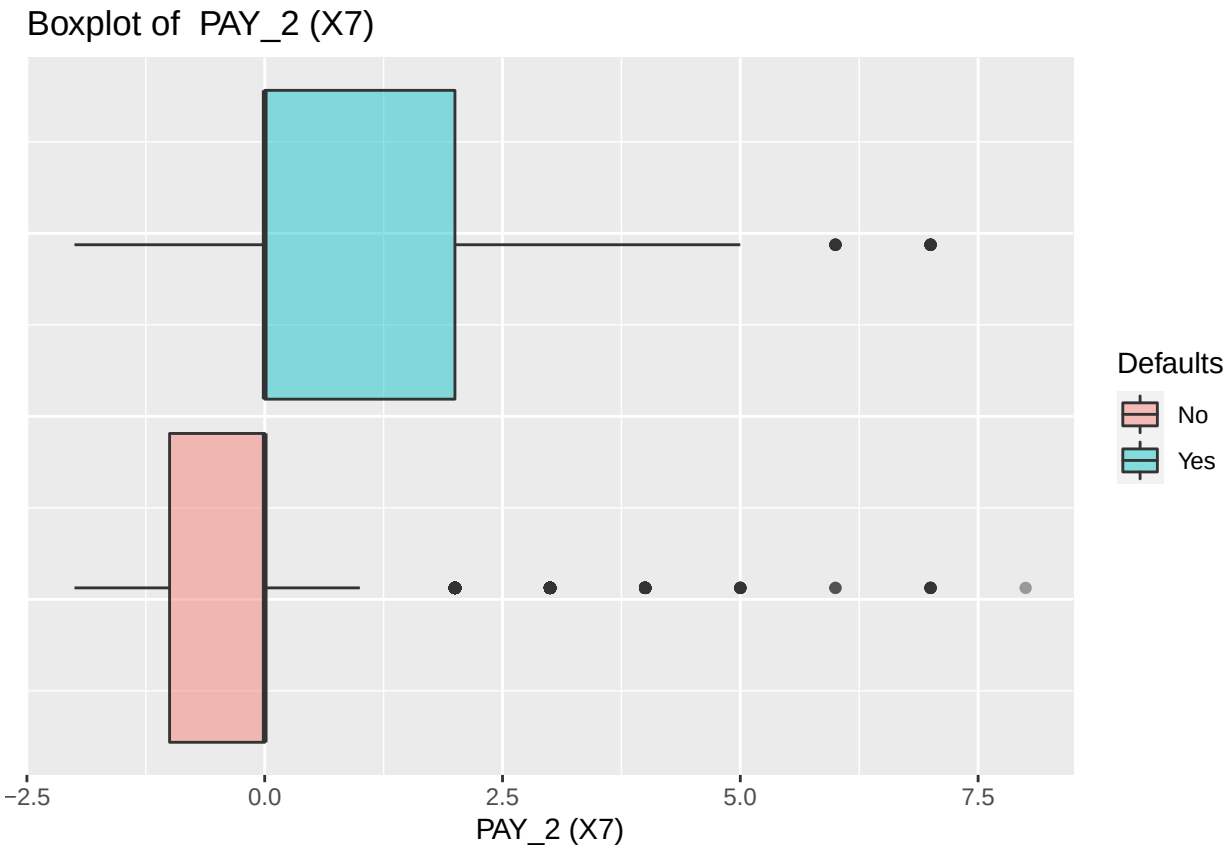
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.1338  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_2
name <- "PAY_2 (X7)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_3 (X8)

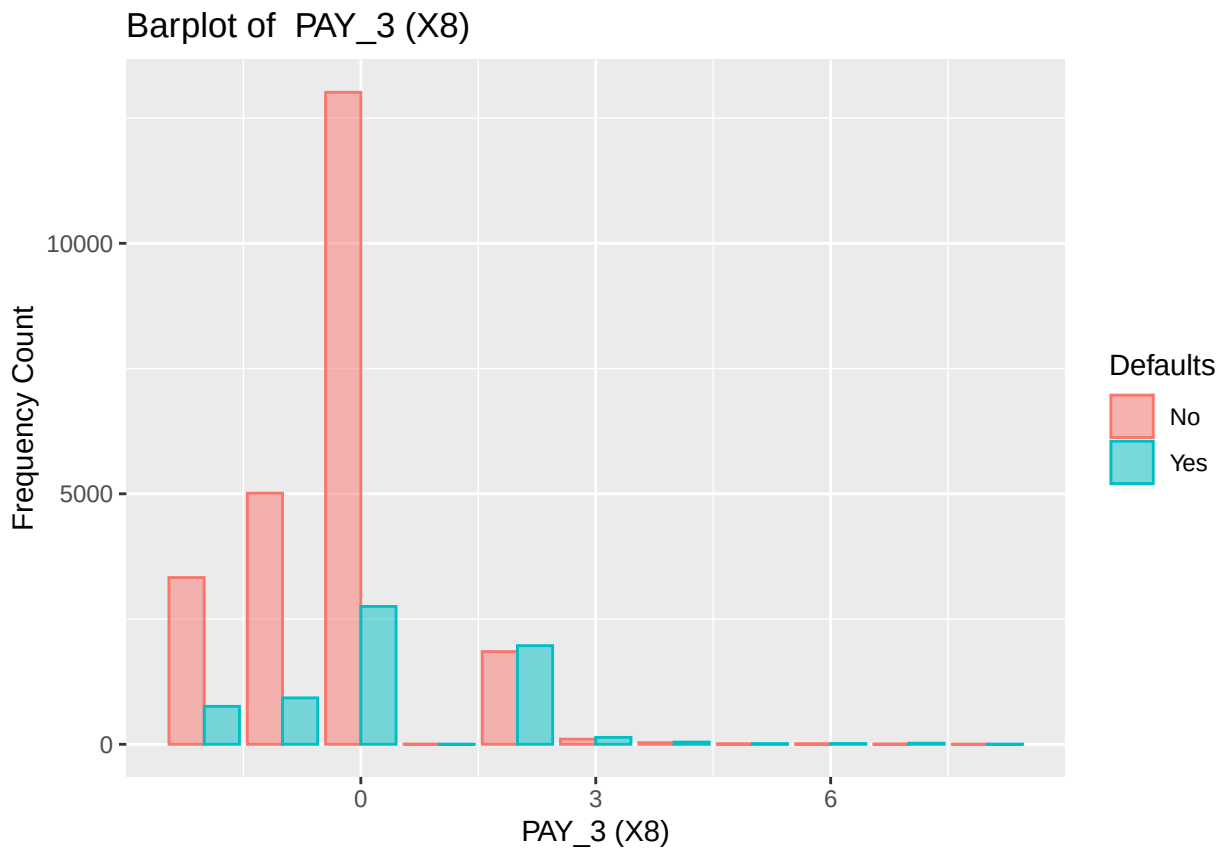
The PAY_3 feature captures the payment status of the individual for the month of **July, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_3)
```

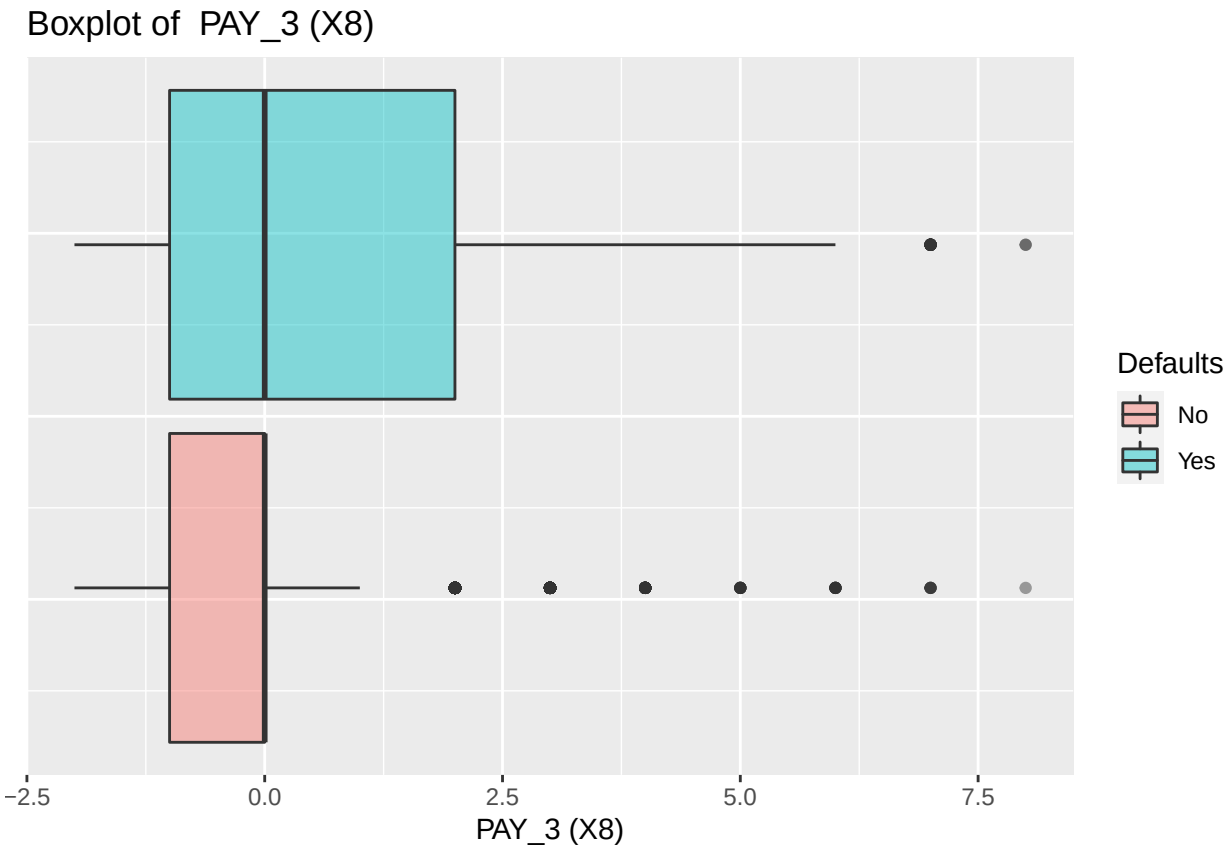
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.1662  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_3
name <- "PAY_3 (X8)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_4 (X9)

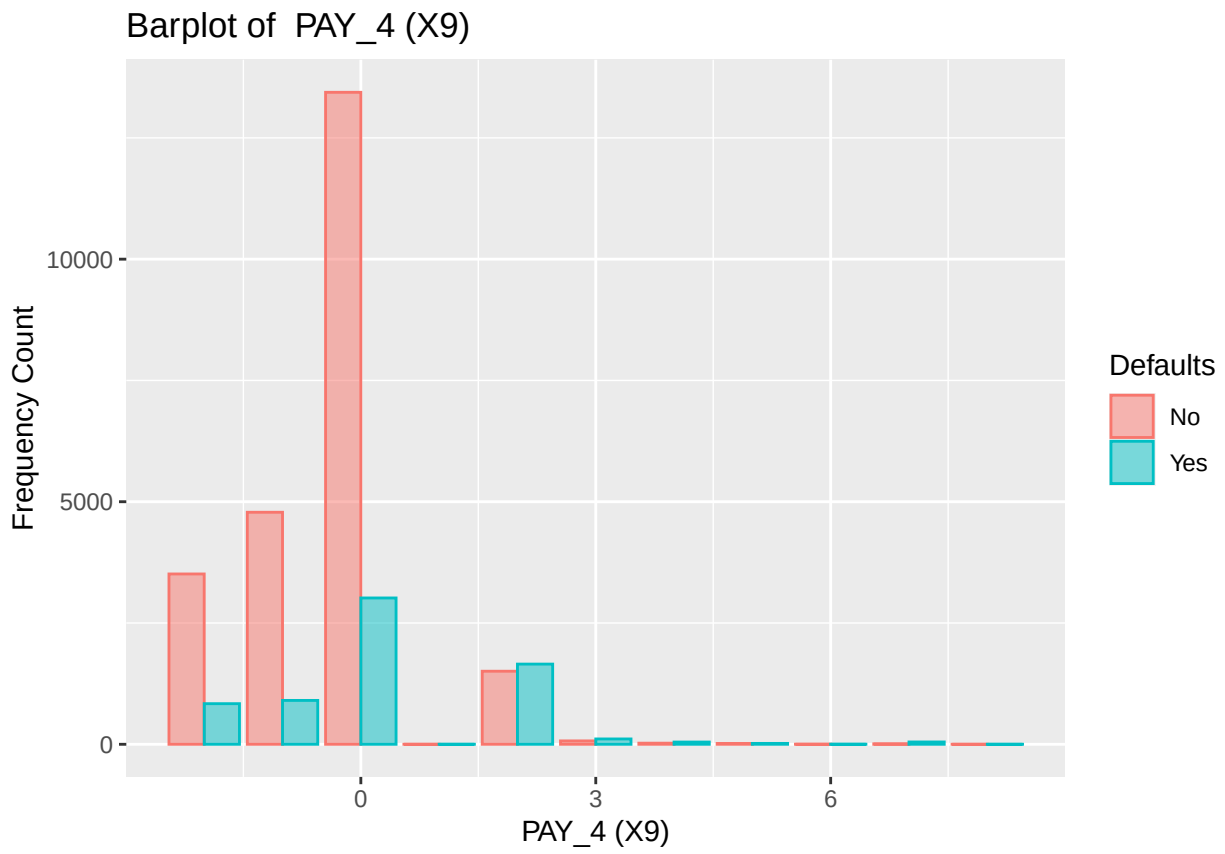
The PAY_4 feature captures the payment status of the individual for the month of **June, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_4)
```

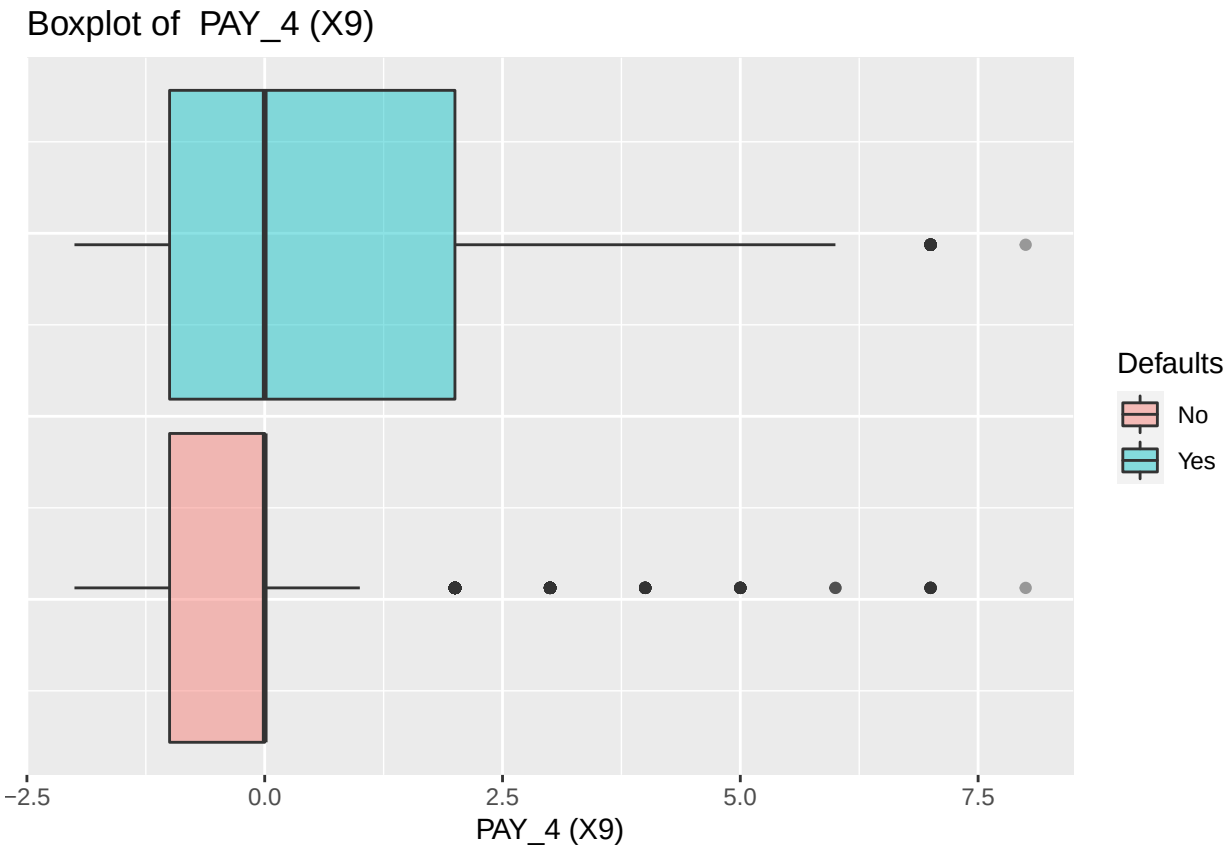
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.2207  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_4
name <- "PAY_4 (X9)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_5 (X10)

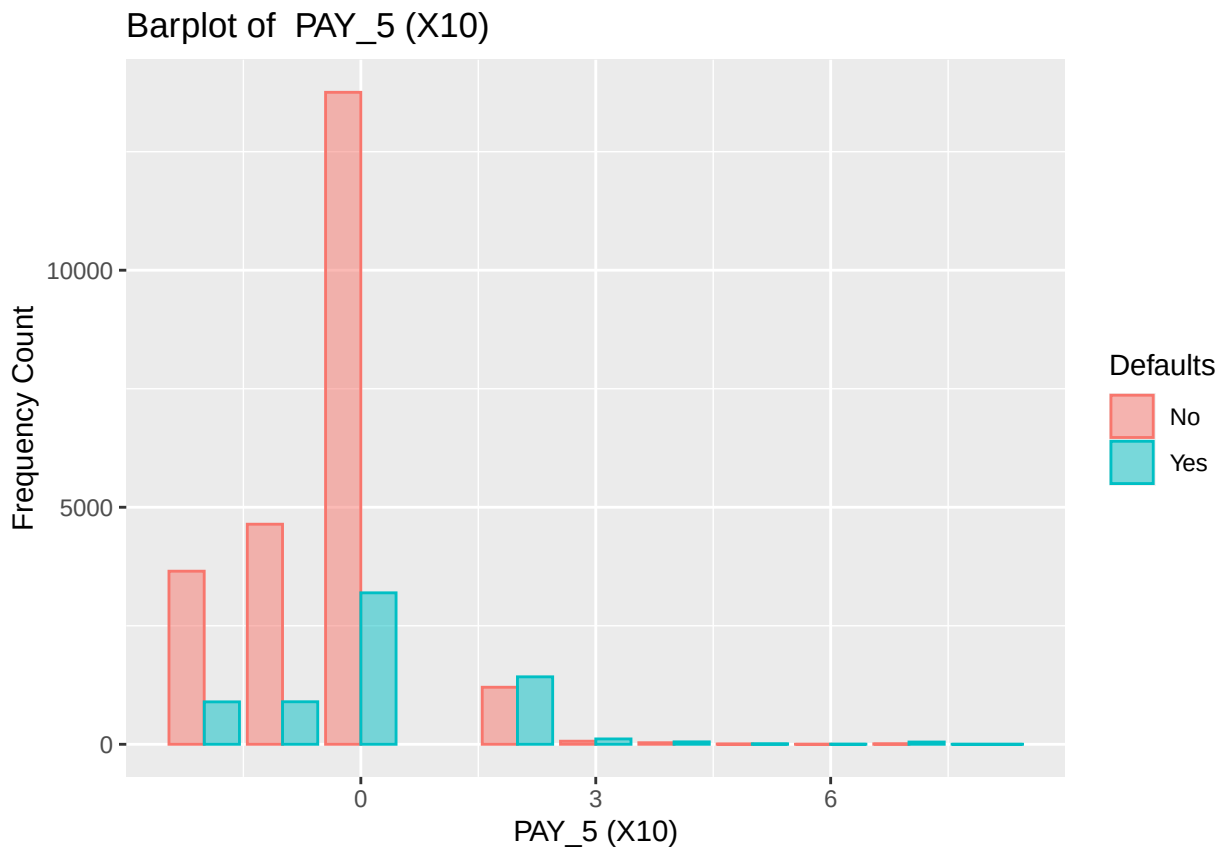
The PAY_5 feature captures the payment status of the individual for the month of **May, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_5)
```

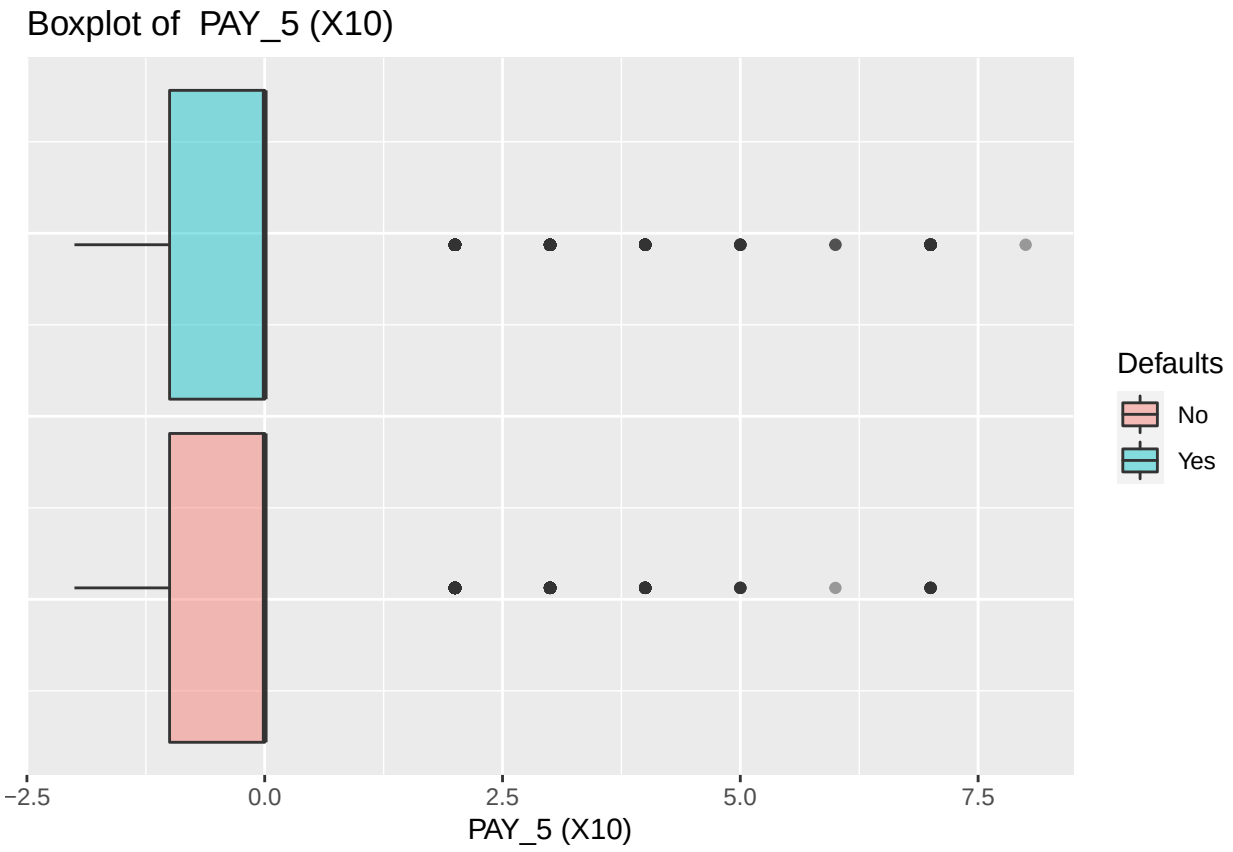
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.2662  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_5
name <- "PAY_5 (X10)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_6 (X11)

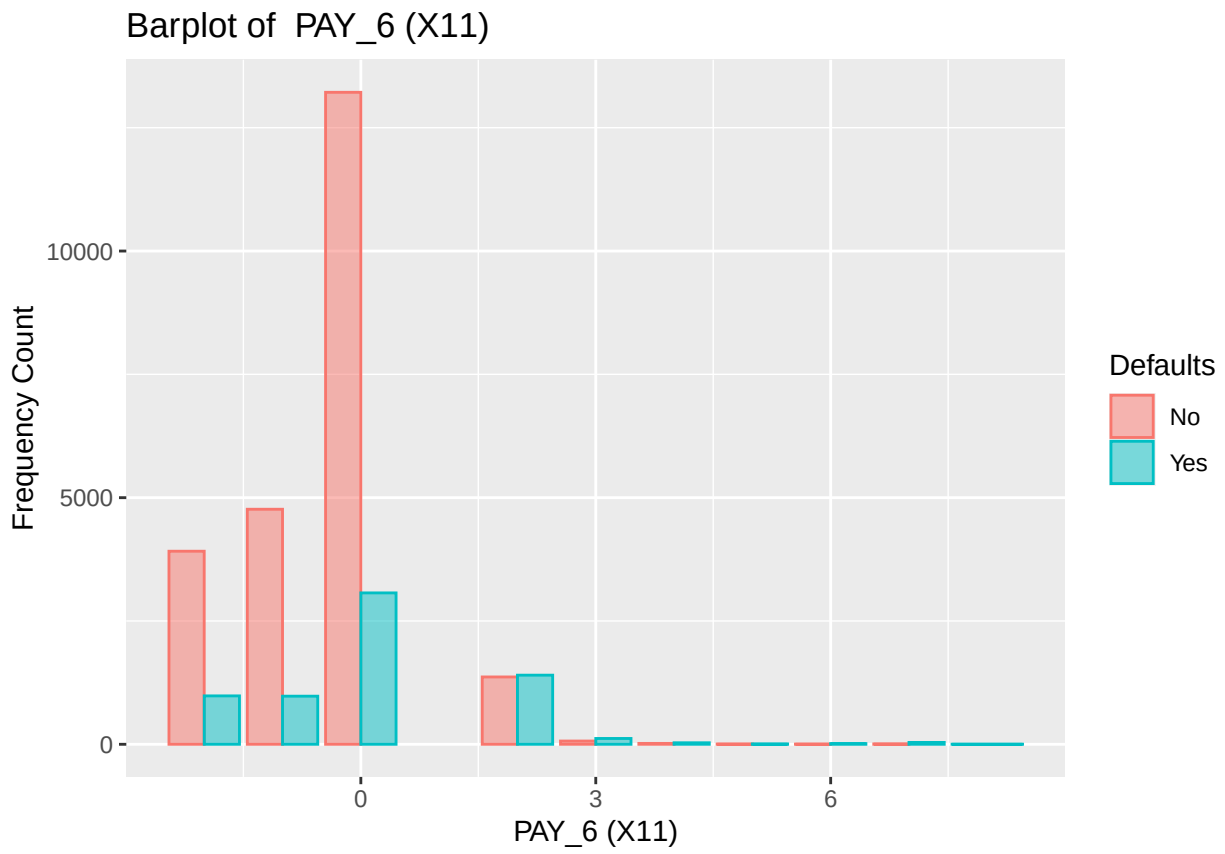
The PAY_6 feature captures the payment status of the individual for the month of **April, 2005**. A negative number represents a duly payment, a 0 represents the month not being paid, and a positive number represents the number of months behind payment.

```
summary(df$PAY_6)
```

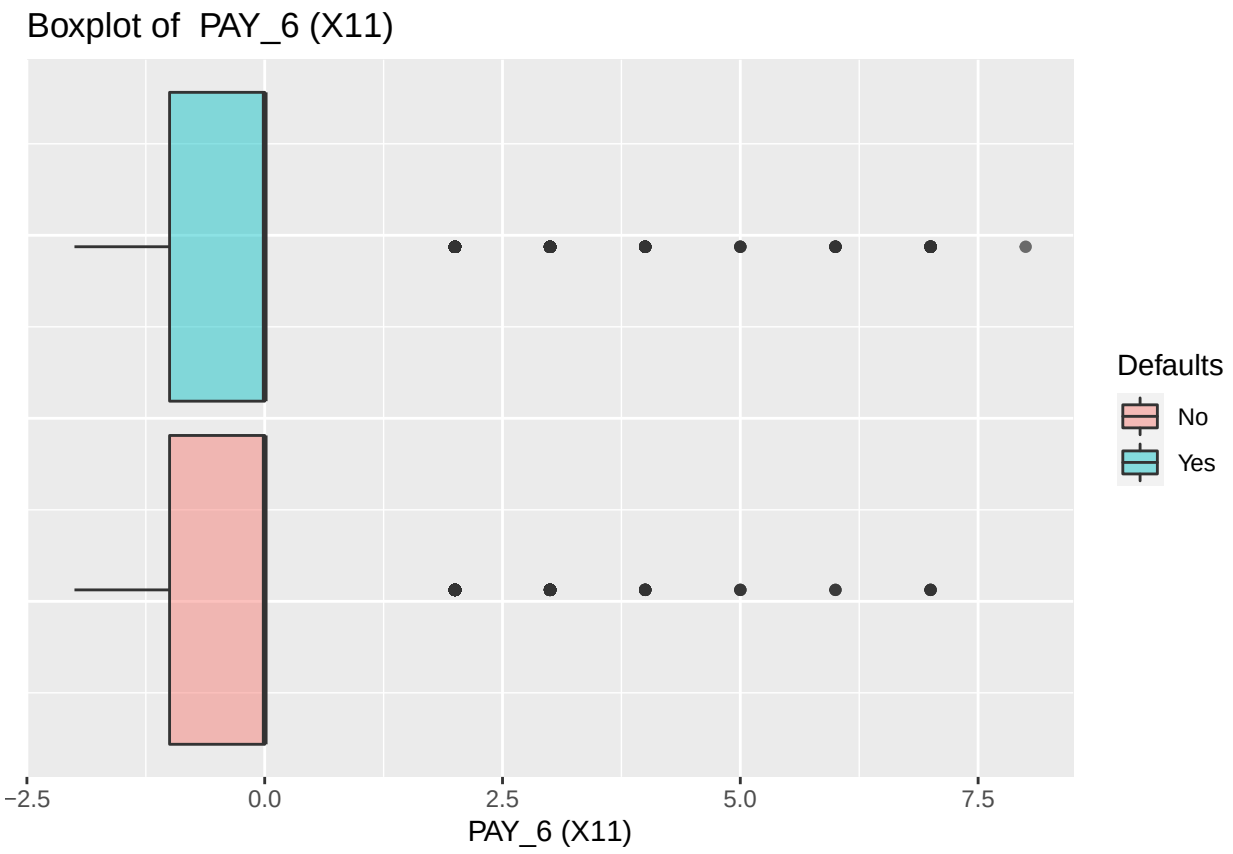
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0000 -1.0000  0.0000 -0.2911  0.0000  8.0000
```

Graphical Analysis

```
x <- df$PAY_6
name <- "PAY_6 (X11)"
bar_comp_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT1 (X12)

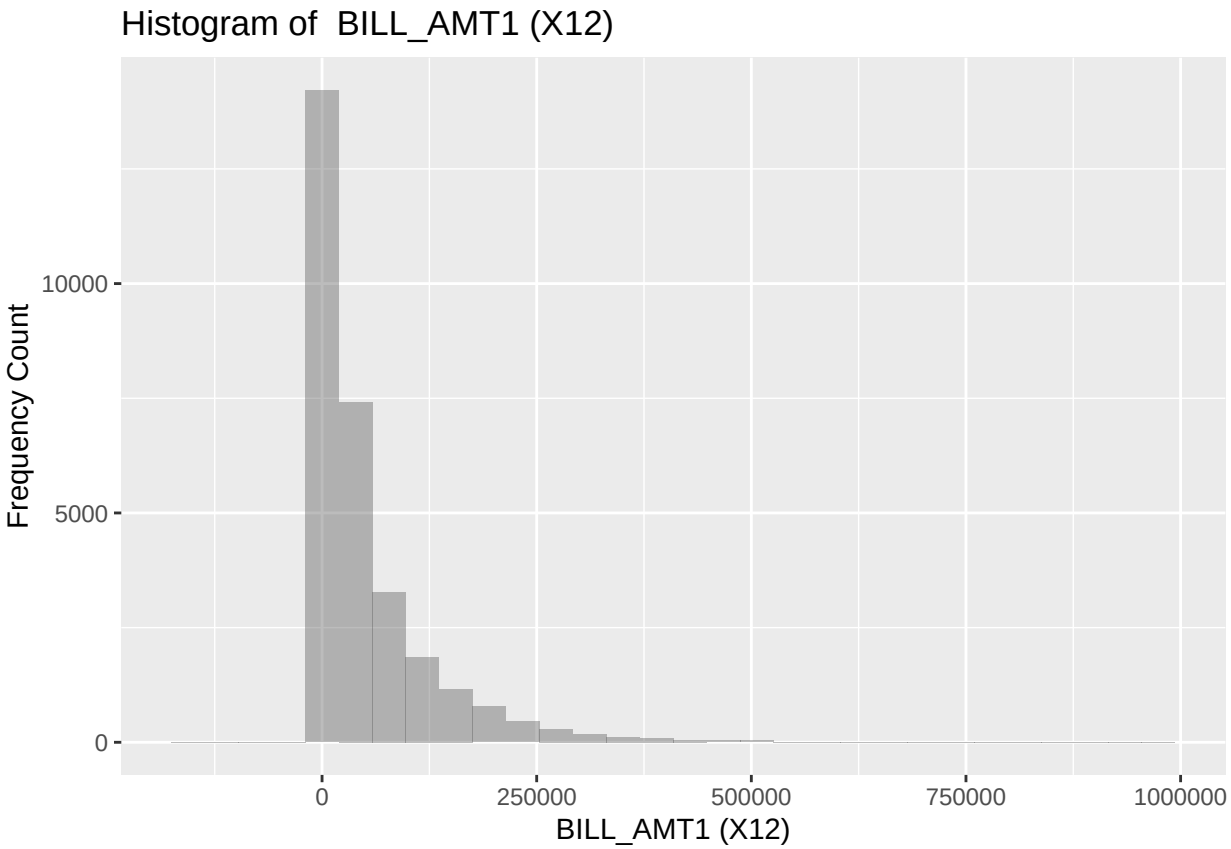
The BILL_AMT1 feature captures the billed amount to the individual for the month of **September, 2005**.

```
summary(df$BILL_AMT1)
```

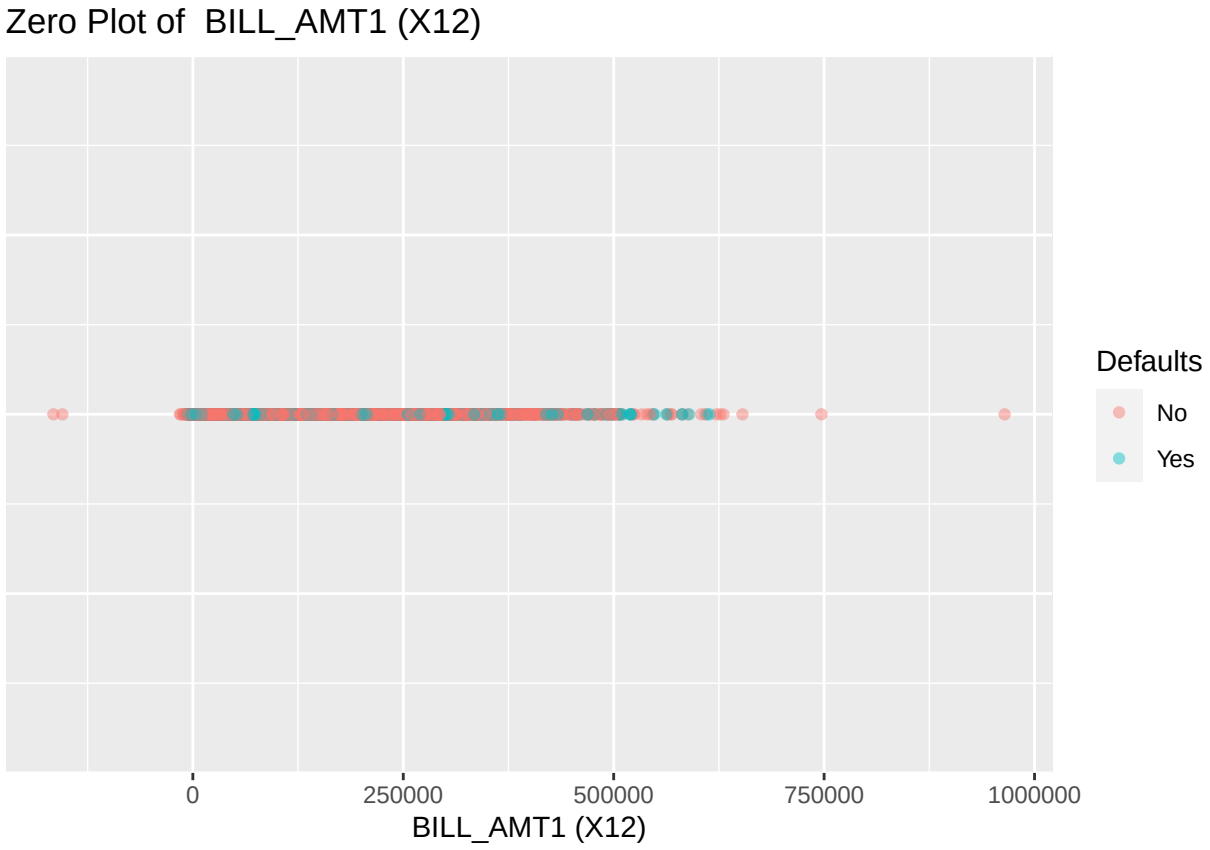
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -165580    3559   22382   51223   67091  964511
```

Graphical Analysis

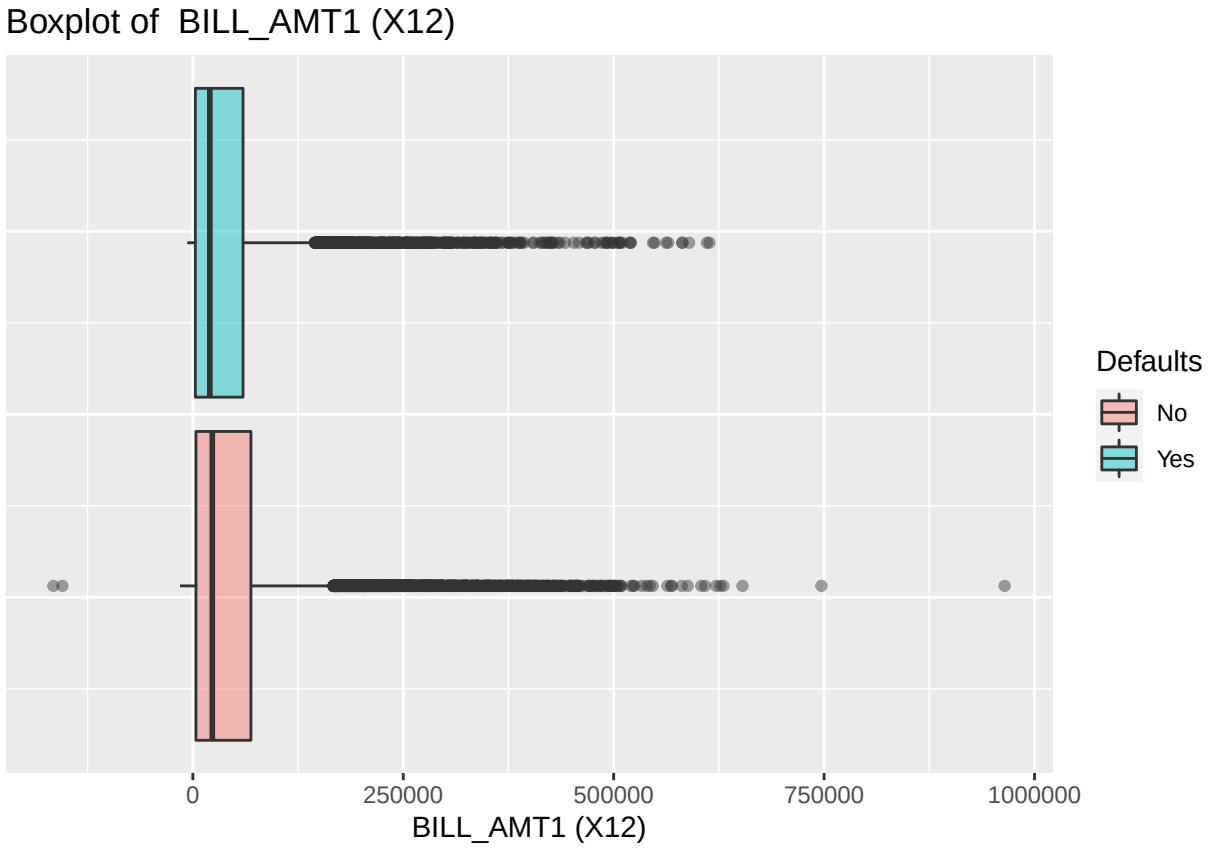
```
x <- df$BILL_AMT1
name <- "BILL_AMT1 (X12)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT2 (X13)

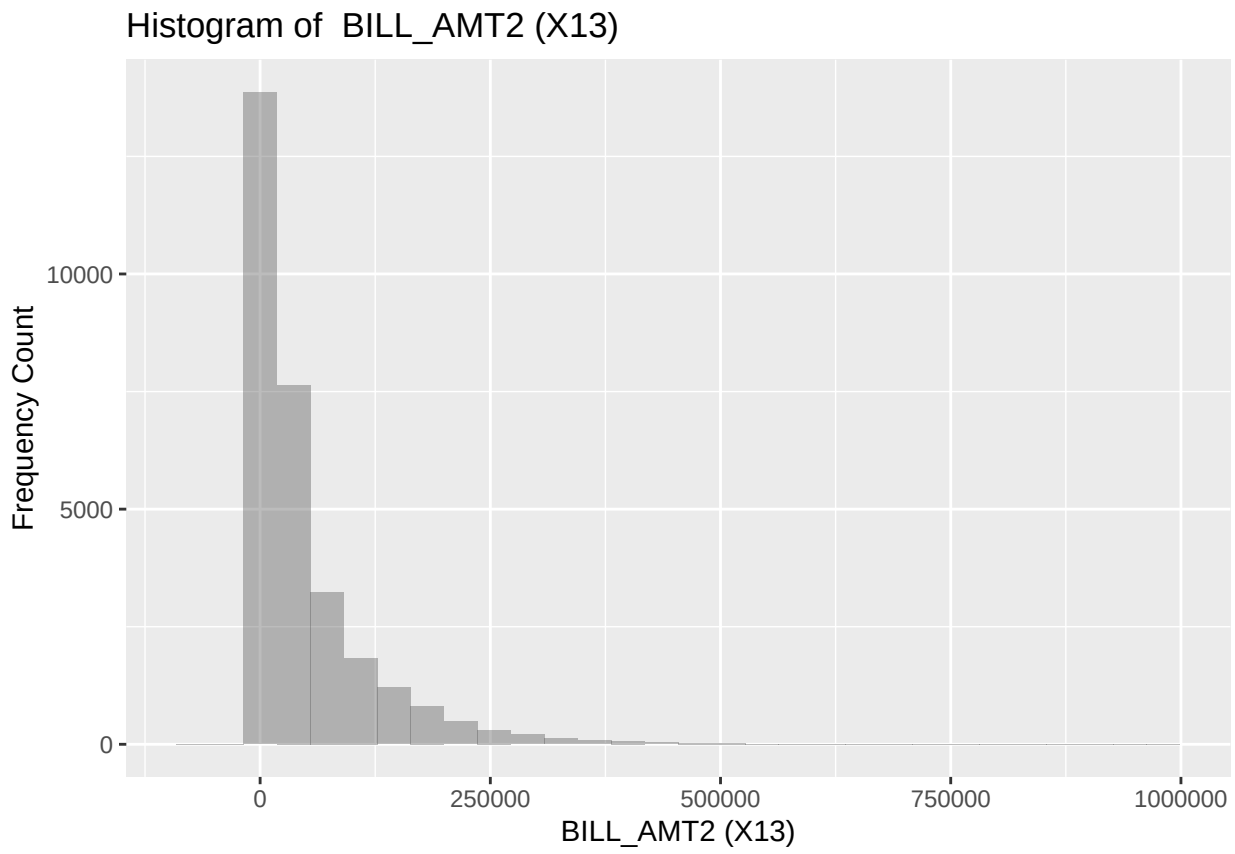
The BILL_AMT2 feature captures the billed amount to the individual for the month of **August, 2005**.

```
summary(df$BILL_AMT2)
```

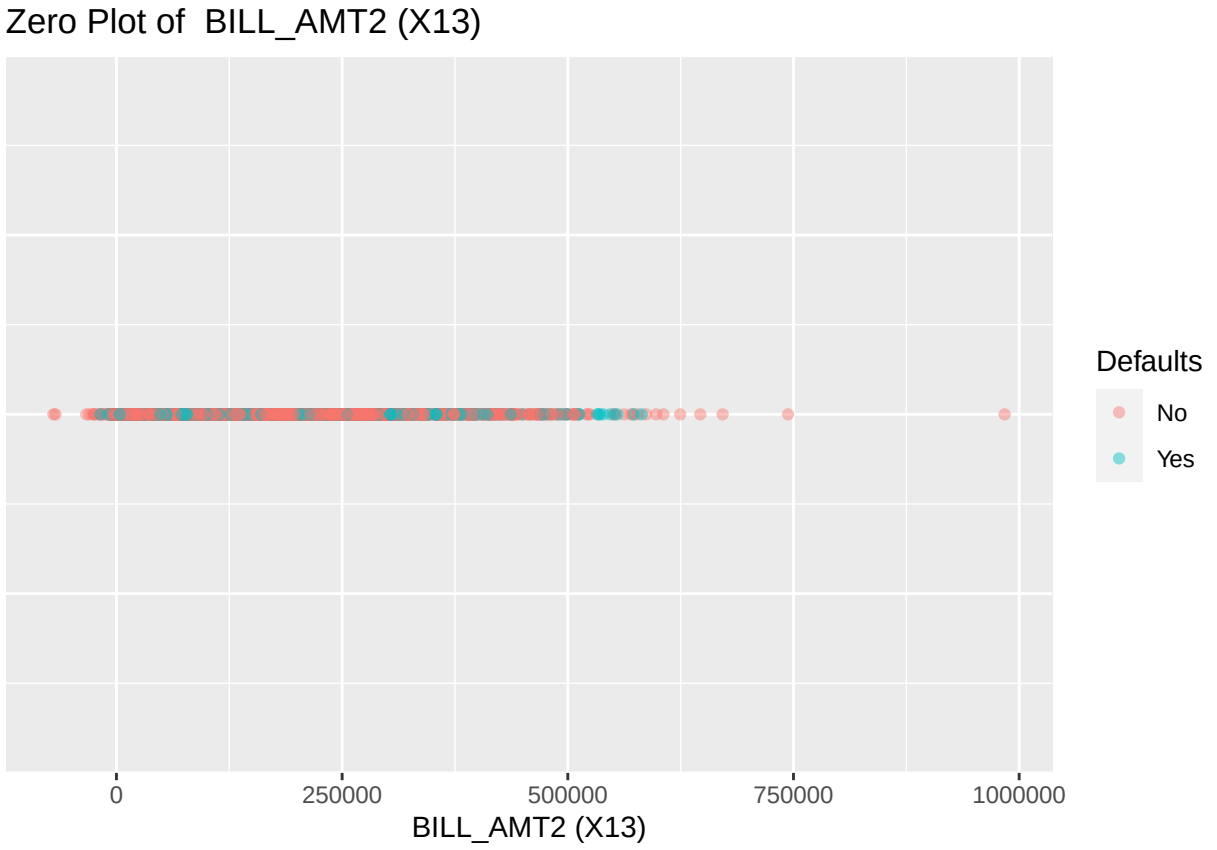
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -69777    2985   21200   49179   64006  983931
```

Graphical Analysis

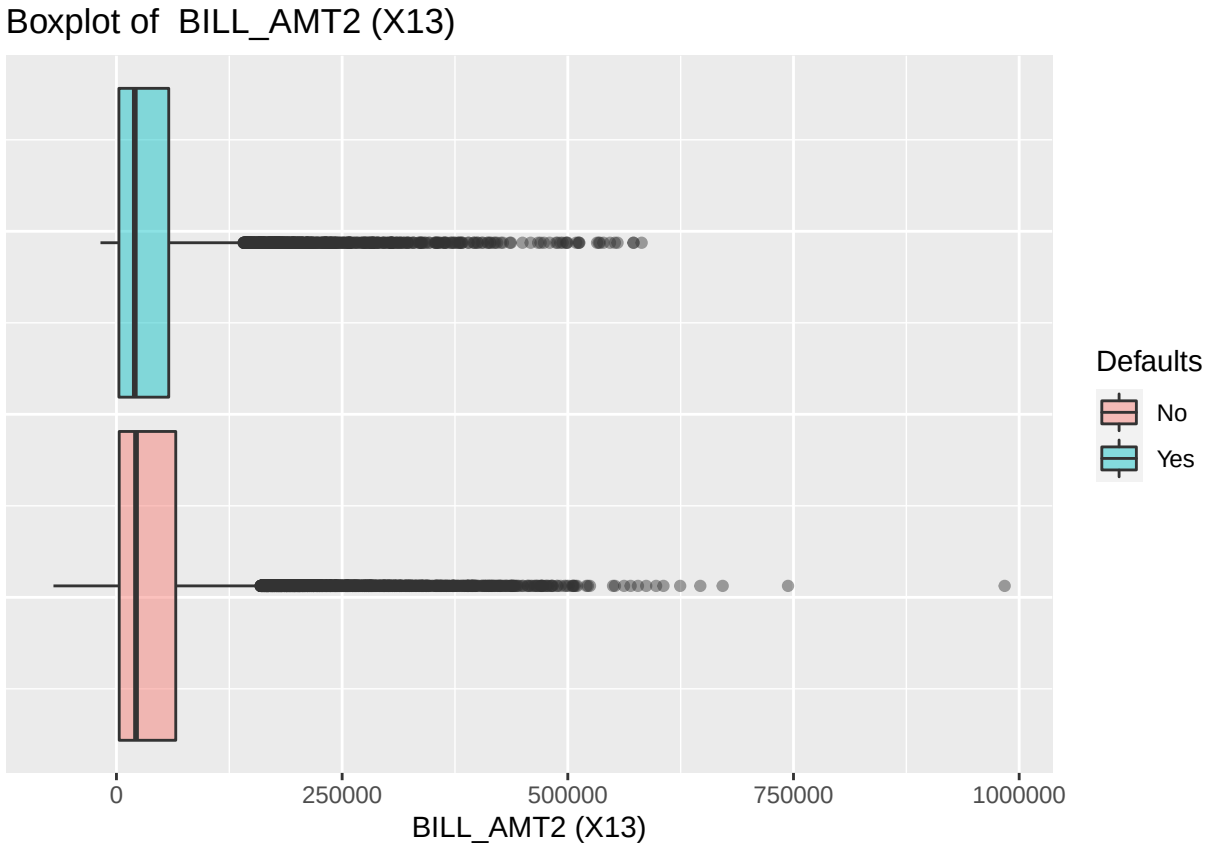
```
x <- df$BILL_AMT2
name <- "BILL_AMT2 (X13)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT3 (X14)

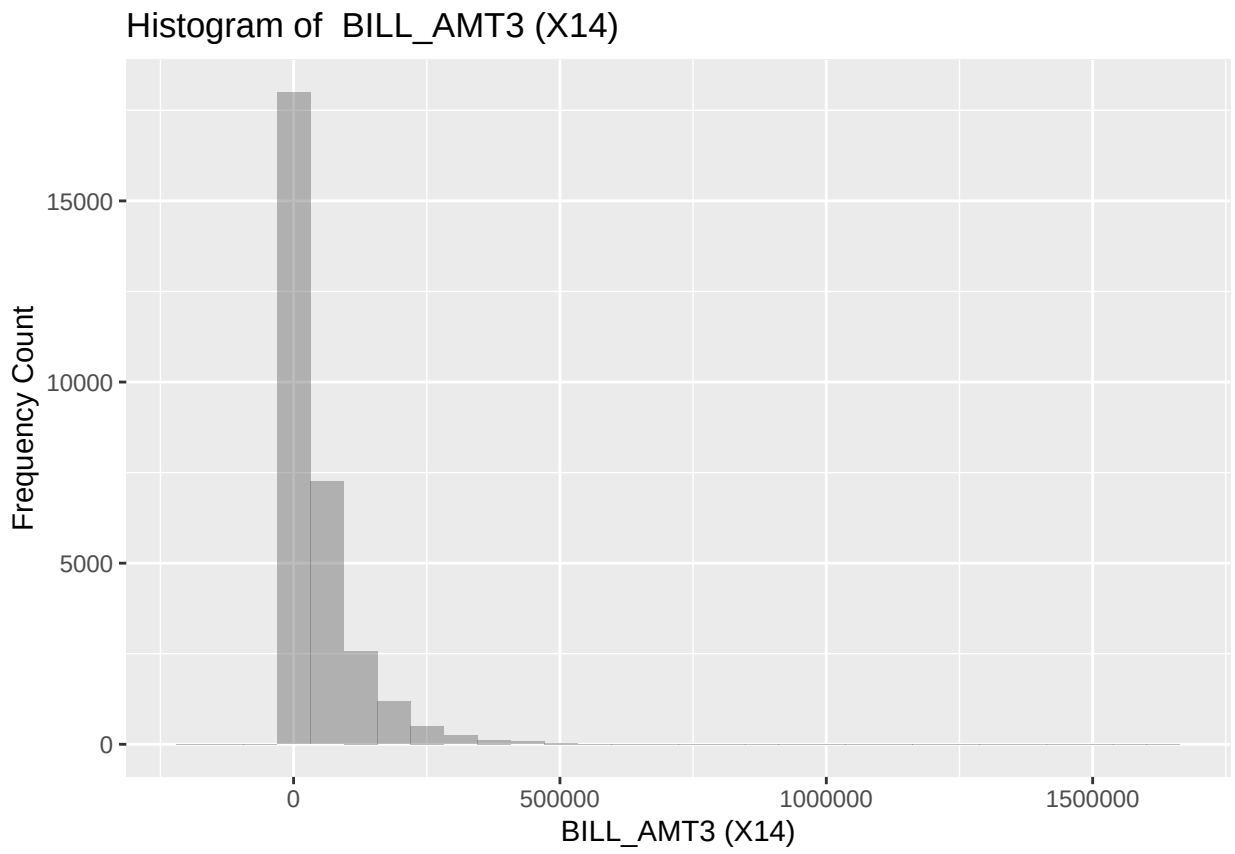
The BILL_AMT3 feature captures the billed amount to the individual for the month of **July, 2005**.

```
summary(df$BILL_AMT3)
```

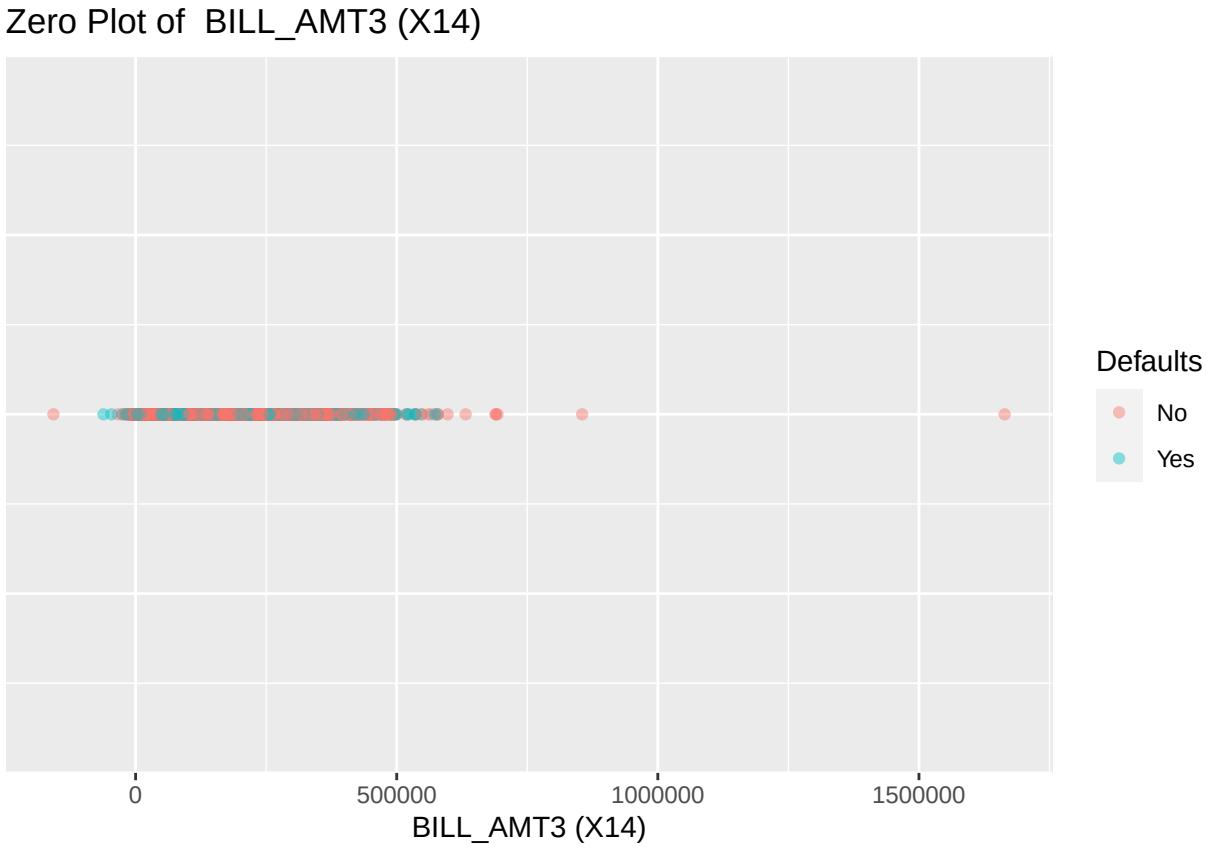
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -157264    2666   20089   47013   60165 1664089
```

Graphical Analysis

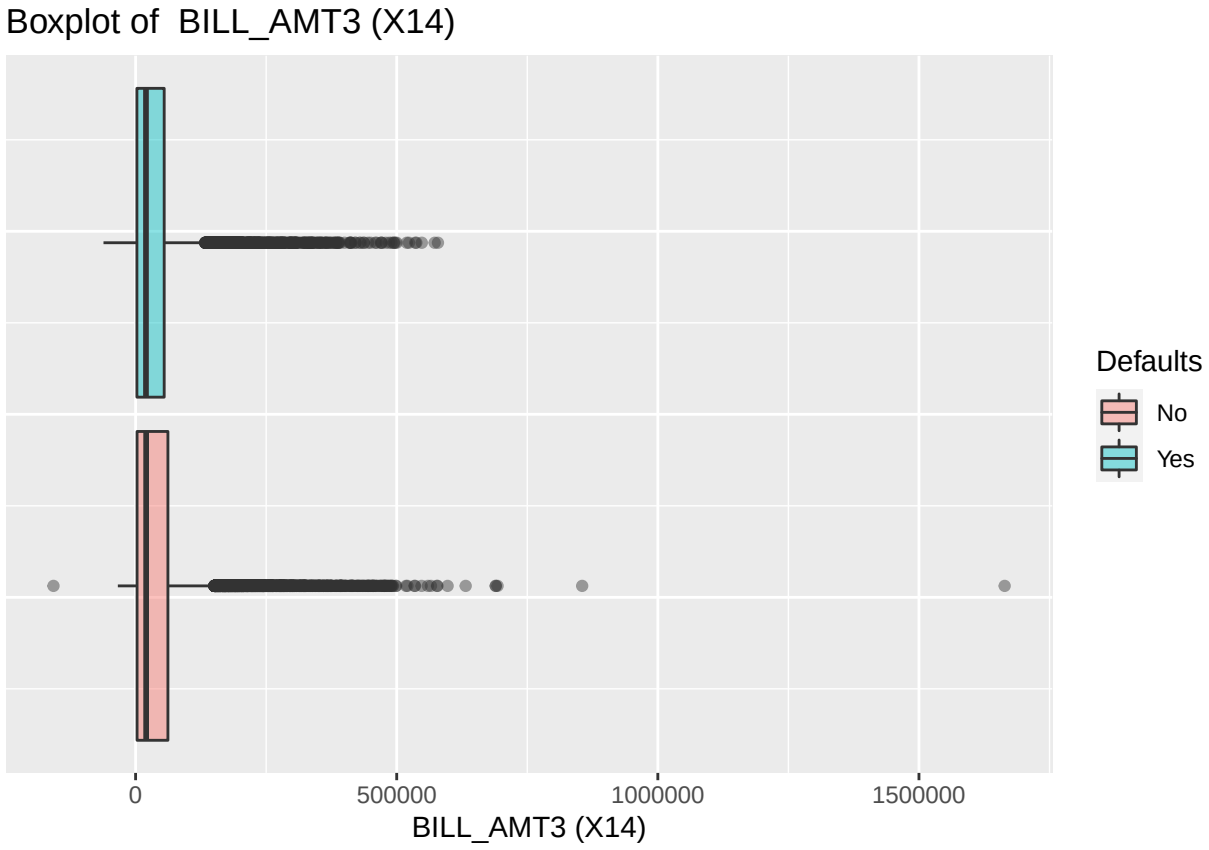
```
x <- df$BILL_AMT3
name <- "BILL_AMT3 (X14)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT4 (X15)

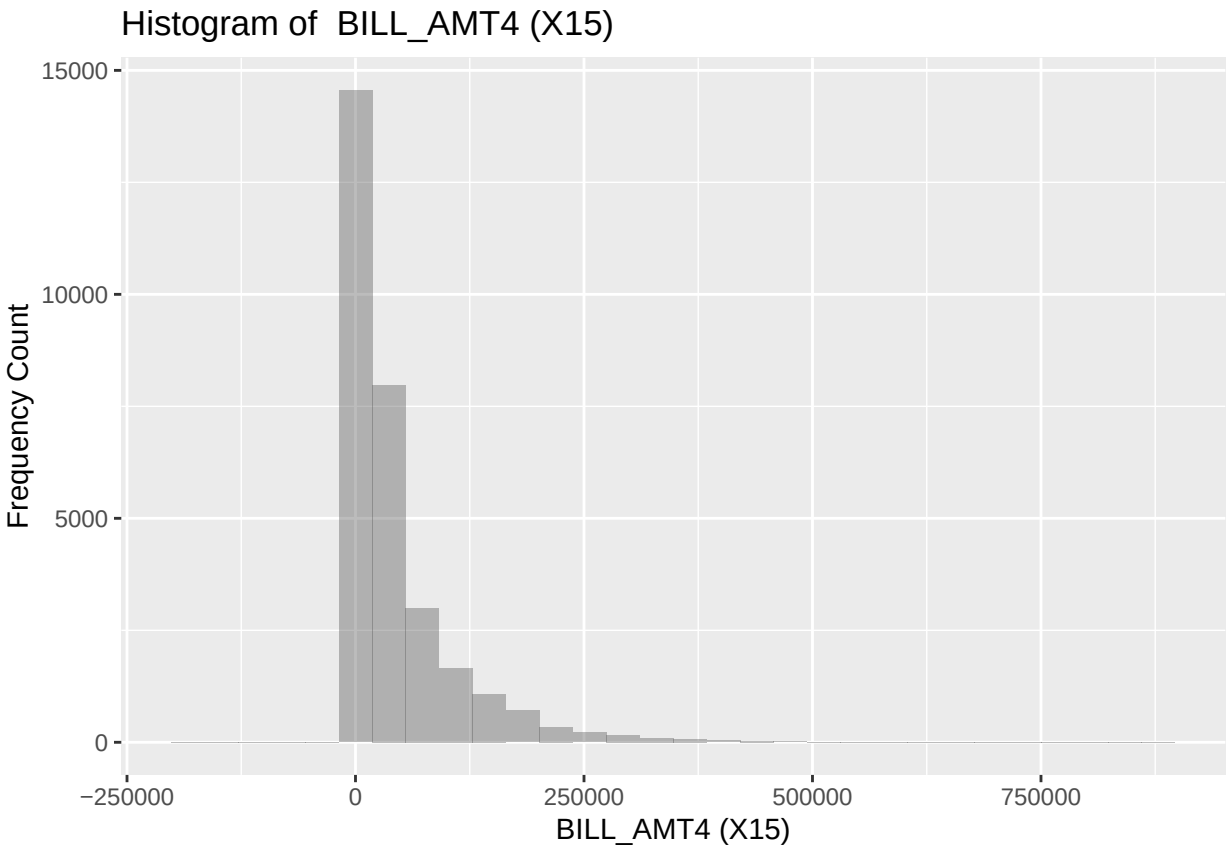
The BILL_AMT4 feature captures the billed amount to the individual for the month of **June, 2005**.

```
summary(df$BILL_AMT4)
```

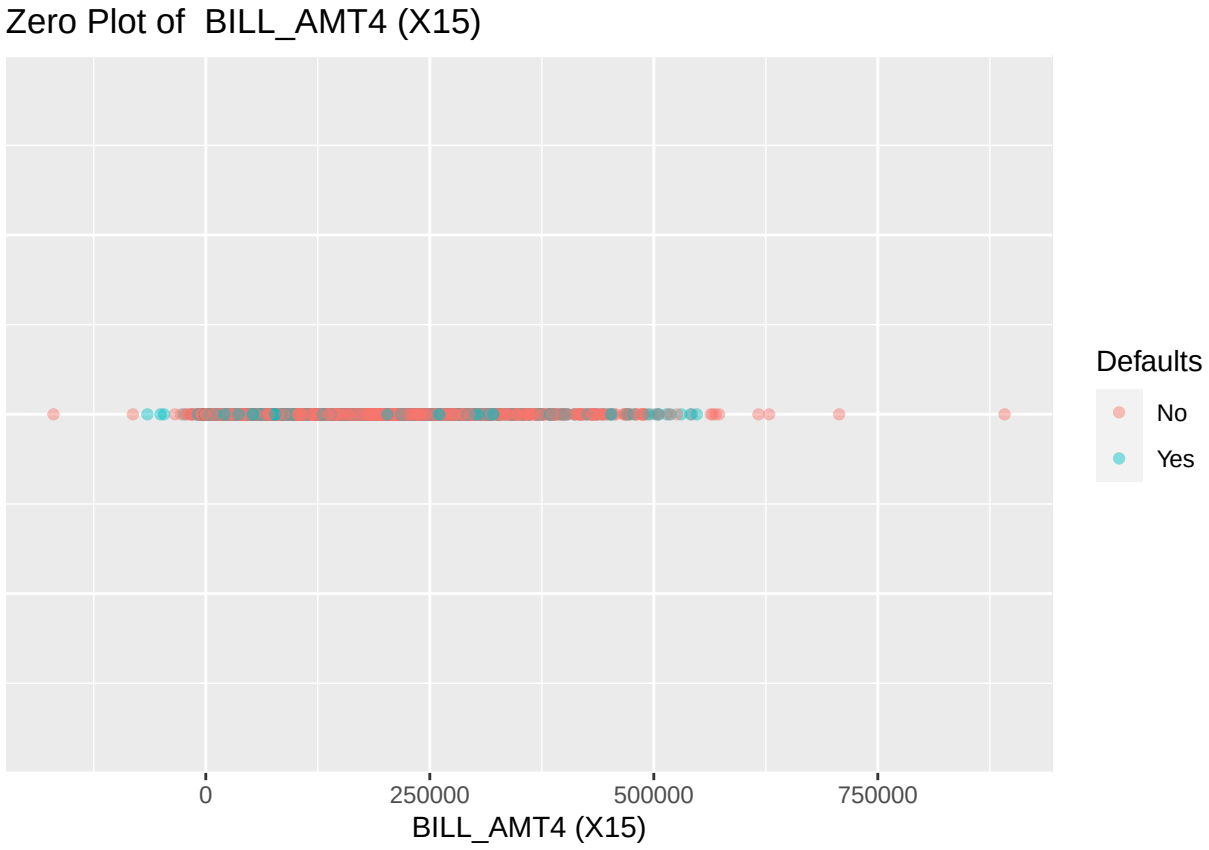
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -170000    2327   19052   43263   54506  891586
```

Graphical Analysis

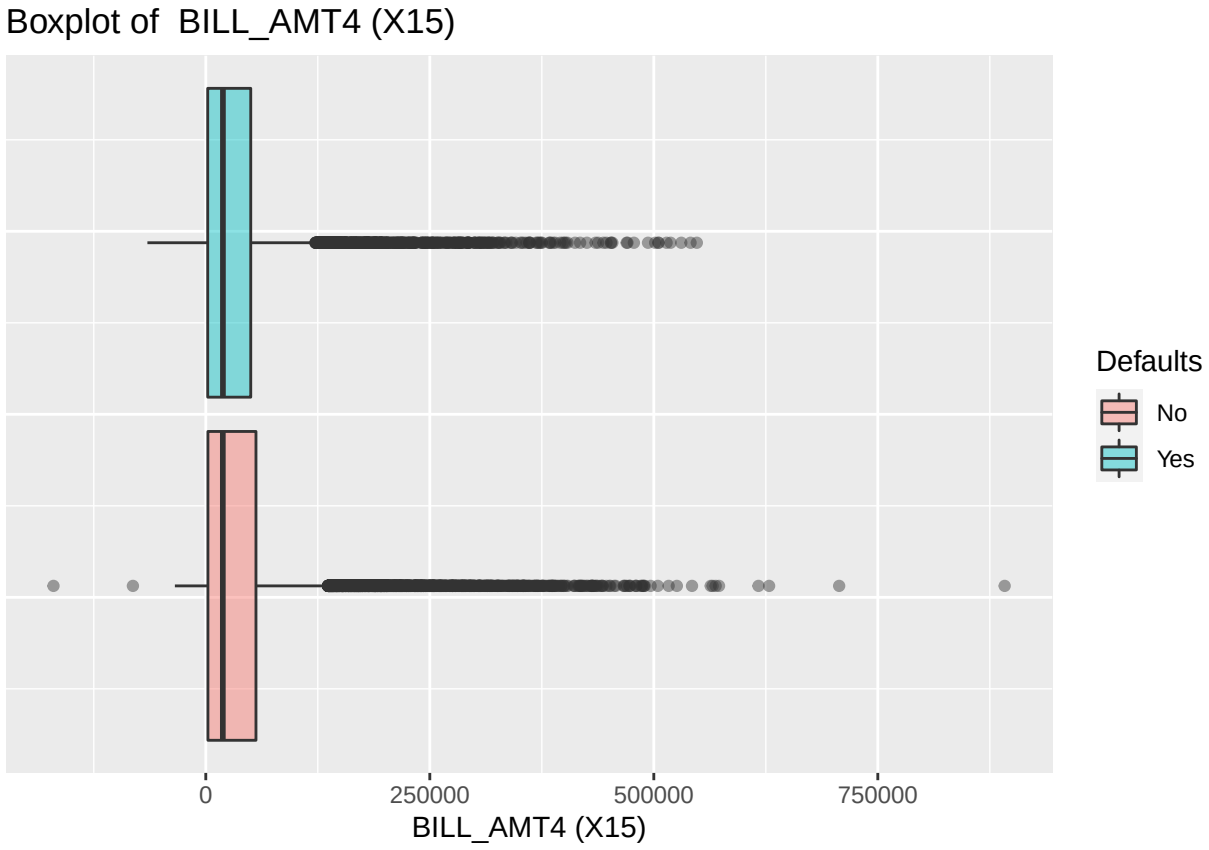
```
x <- df$BILL_AMT4
name <- "BILL_AMT4 (X15)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT5 (X16)

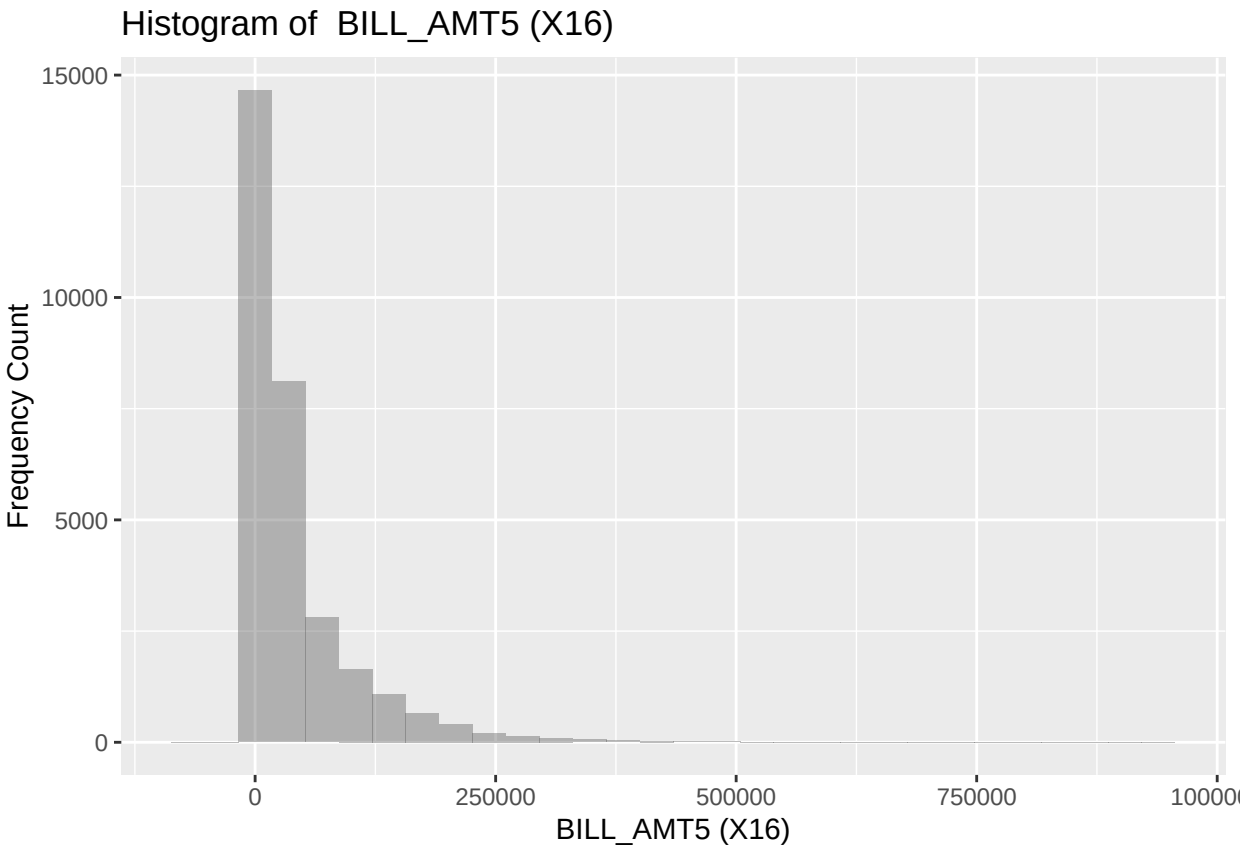
The BILL_AMT5 feature captures the billed amount to the individual for the month of **May, 2005**.

```
summary(df$BILL_AMT5)
```

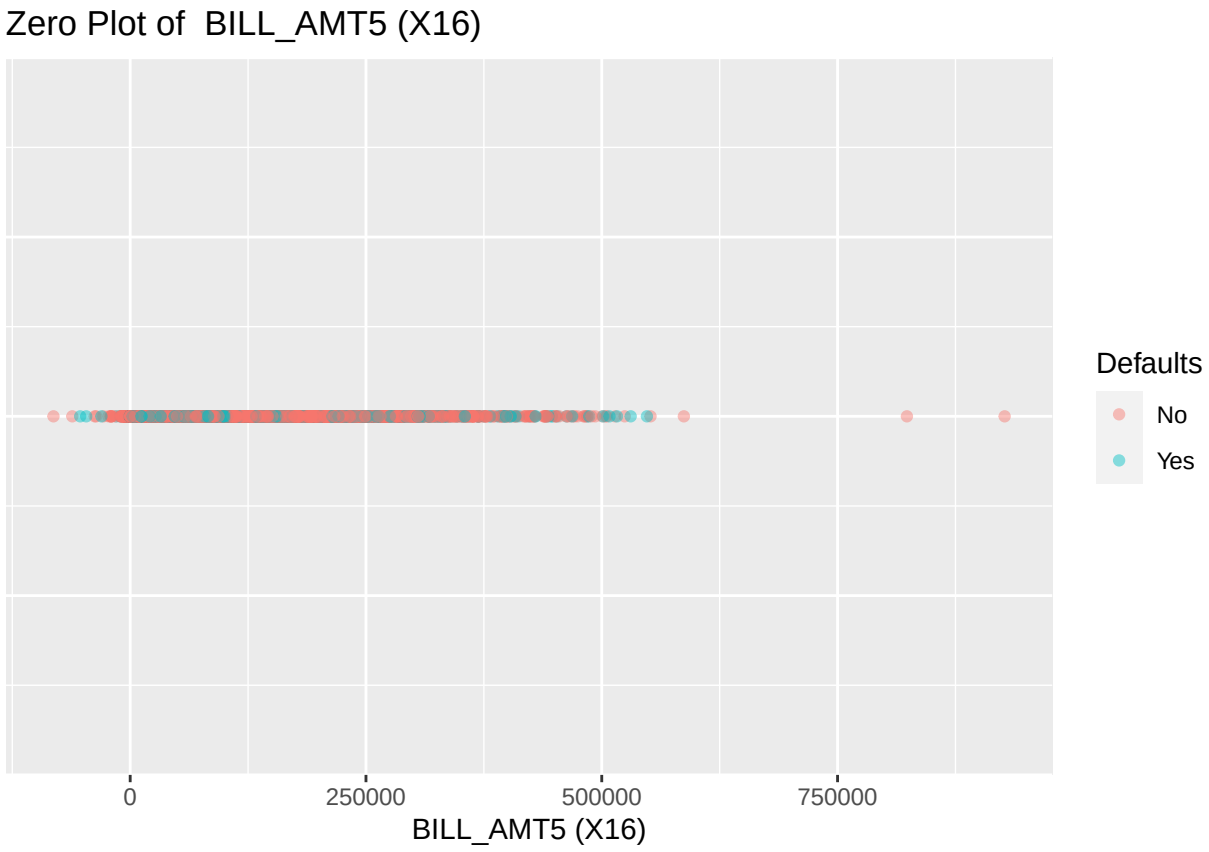
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -81334    1763   18105   40311   50191  927171
```

Graphical Analysis

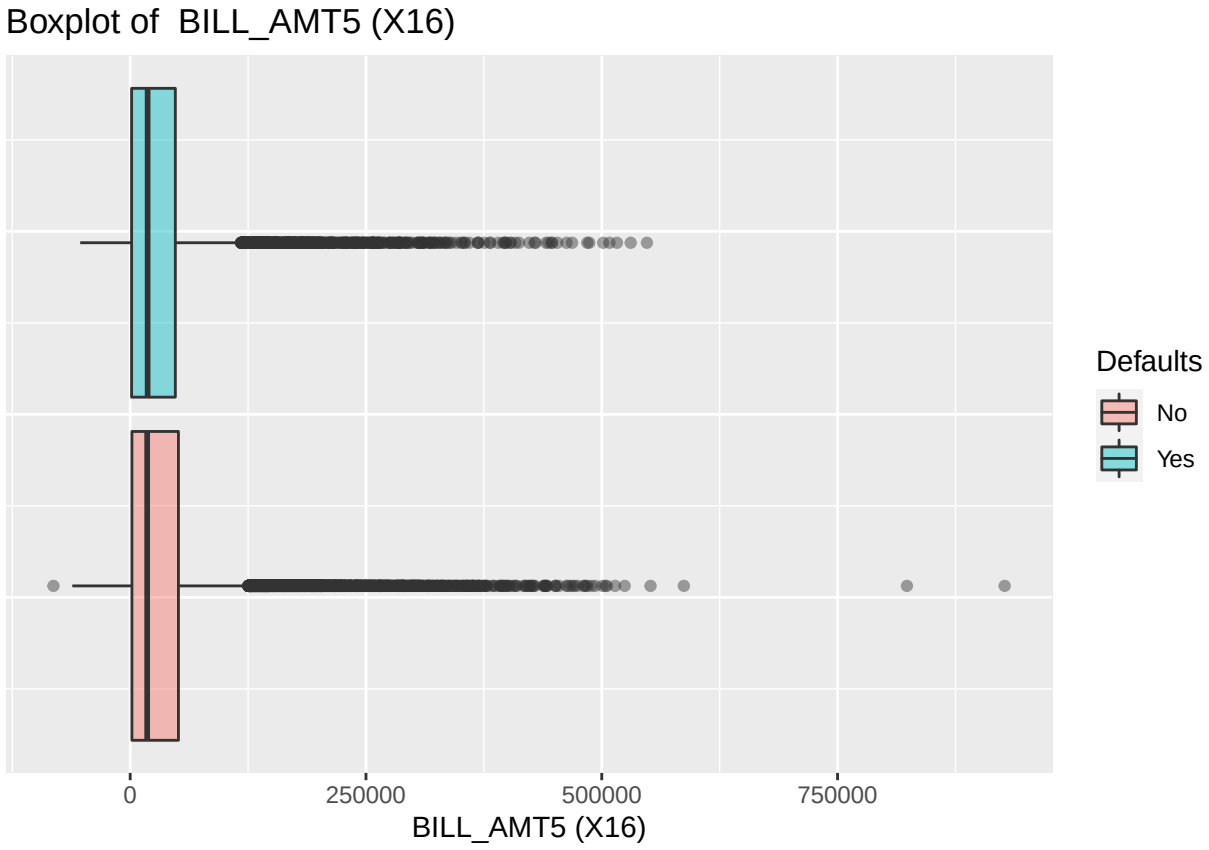
```
x <- df$BILL_AMT5
name <- "BILL_AMT5 (X16)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



BILL_AMT6 (X17)

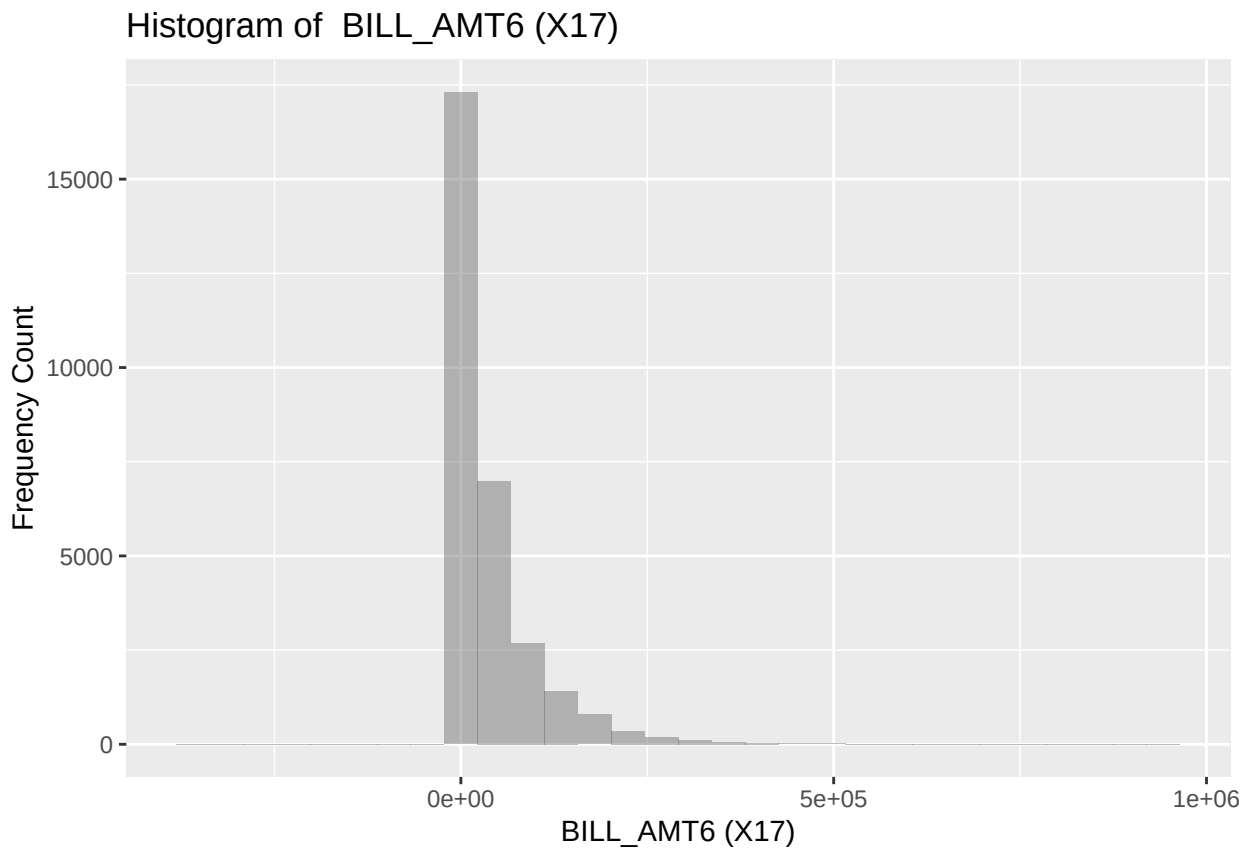
The BILL_AMT6 feature captures the billed amount to the individual for the month of **April, 2005**.

```
summary(df$BILL_AMT6)
```

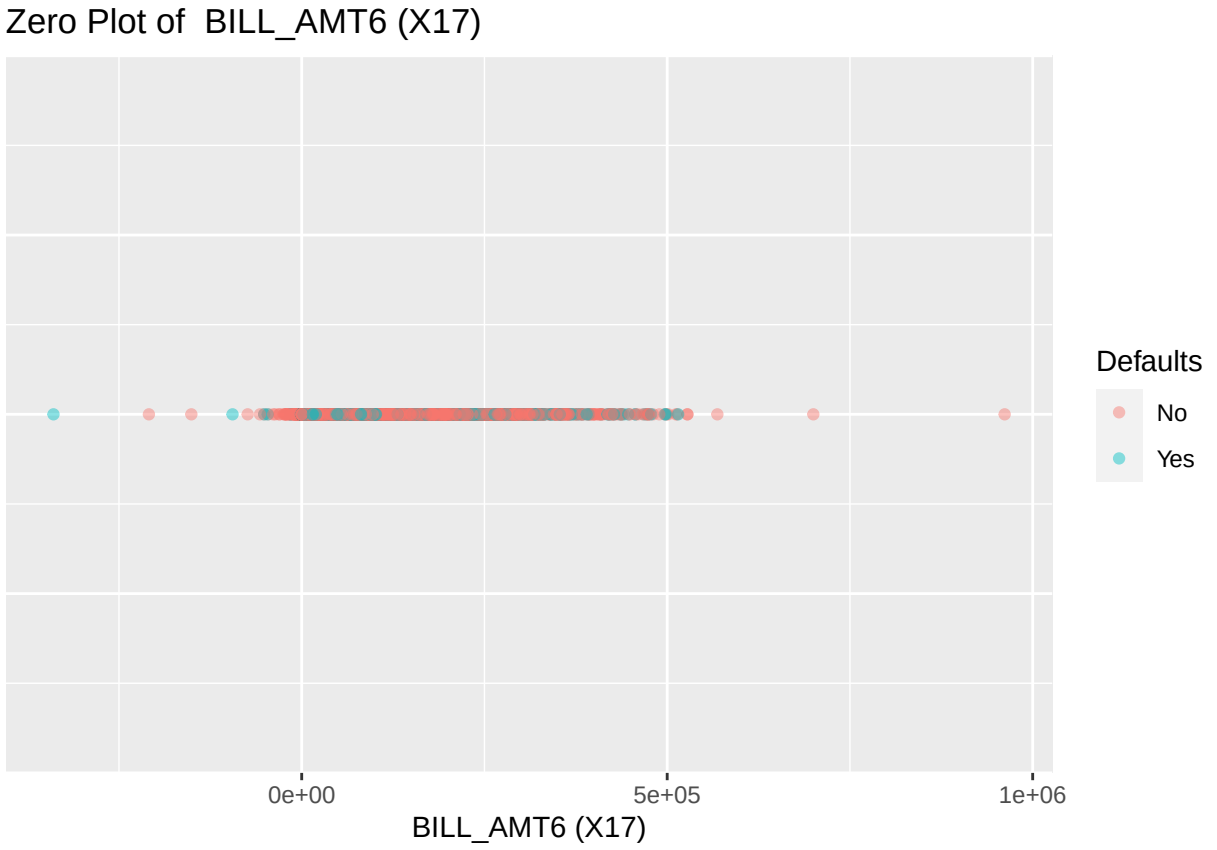
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -339603    1256   17071   38872   49198  961664
```

Graphical Analysis

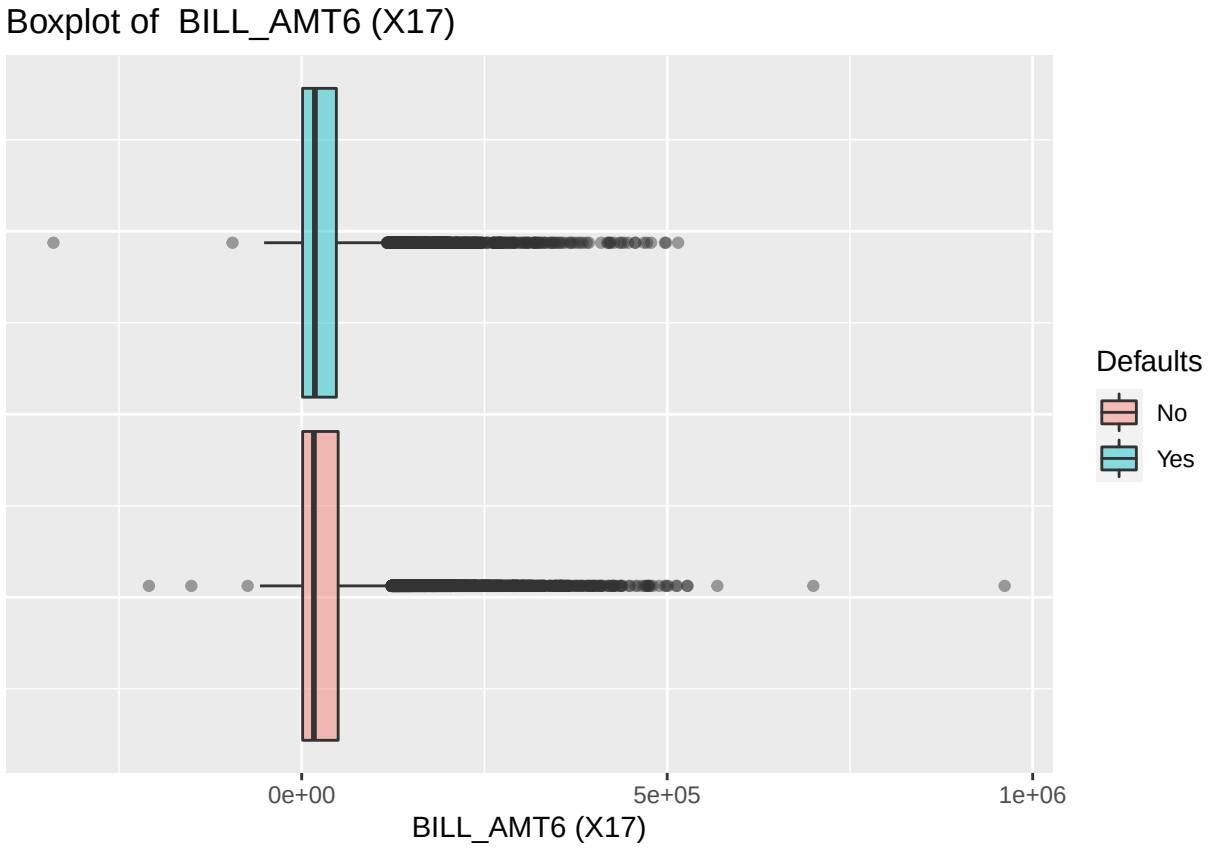
```
x <- df$BILL_AMT6
name <- "BILL_AMT6 (X17)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT1 (X18)

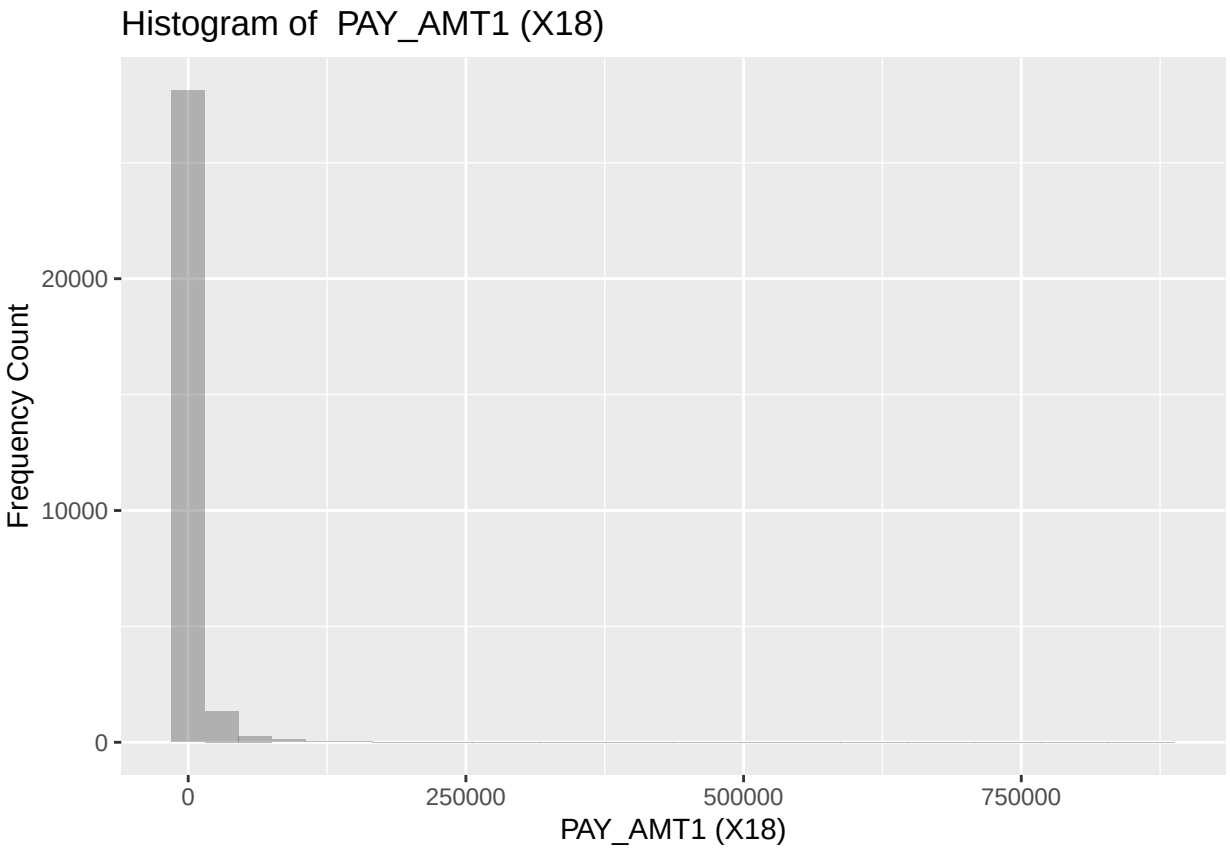
The PAY_AMT1 feature captures the payment amount from the individual for the month of **September, 2005**.

```
summary(df$PAY_AMT1)
```

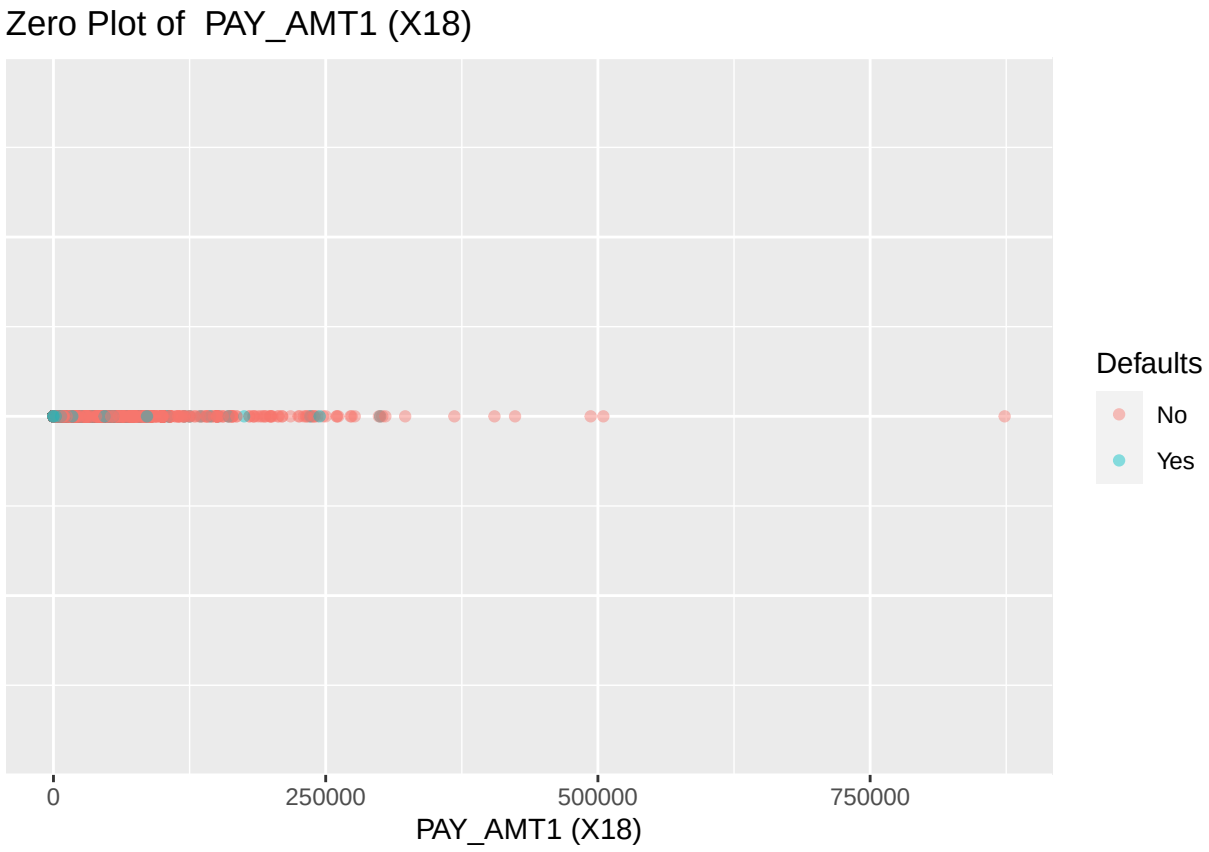
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    1000    2100    5664    5006  873552
```

Graphical Analysis

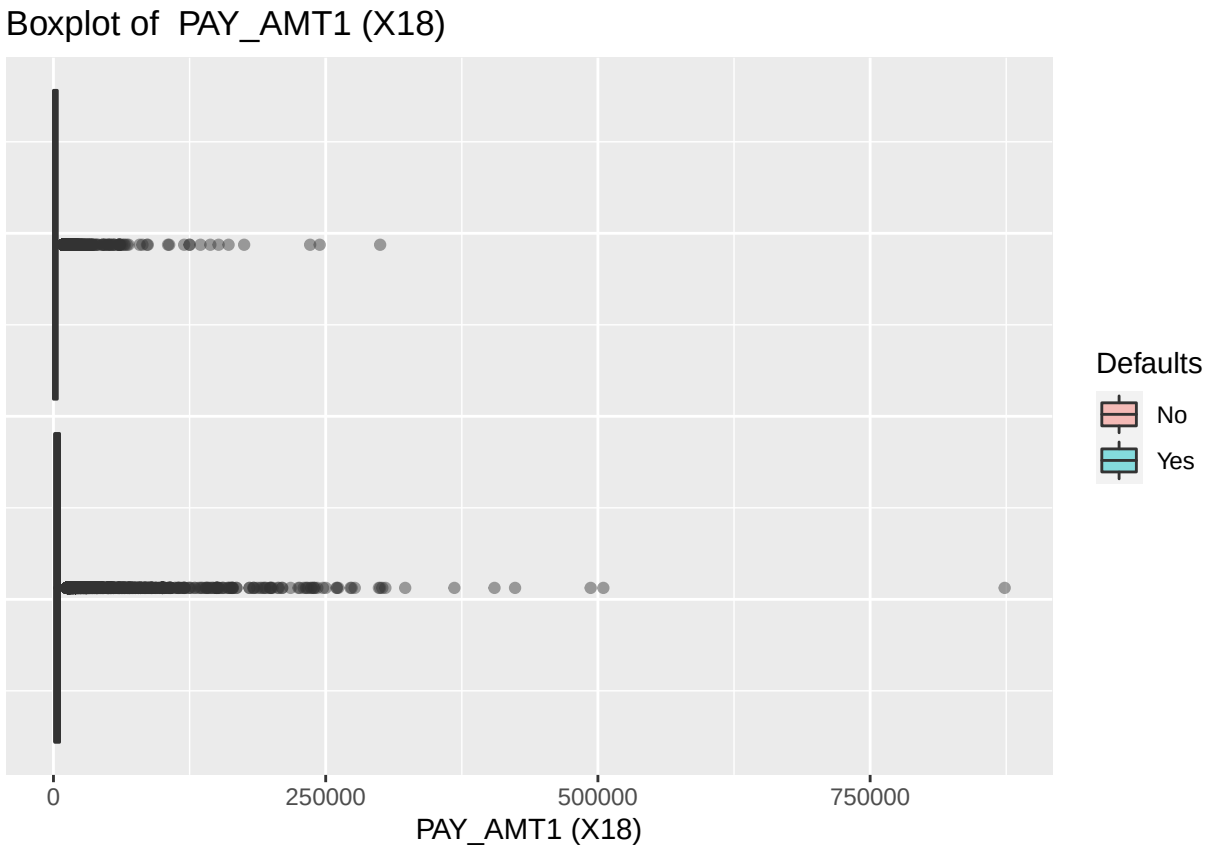
```
x <- df$PAY_AMT1
name <- "PAY_AMT1 (X18)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT2 (X19)

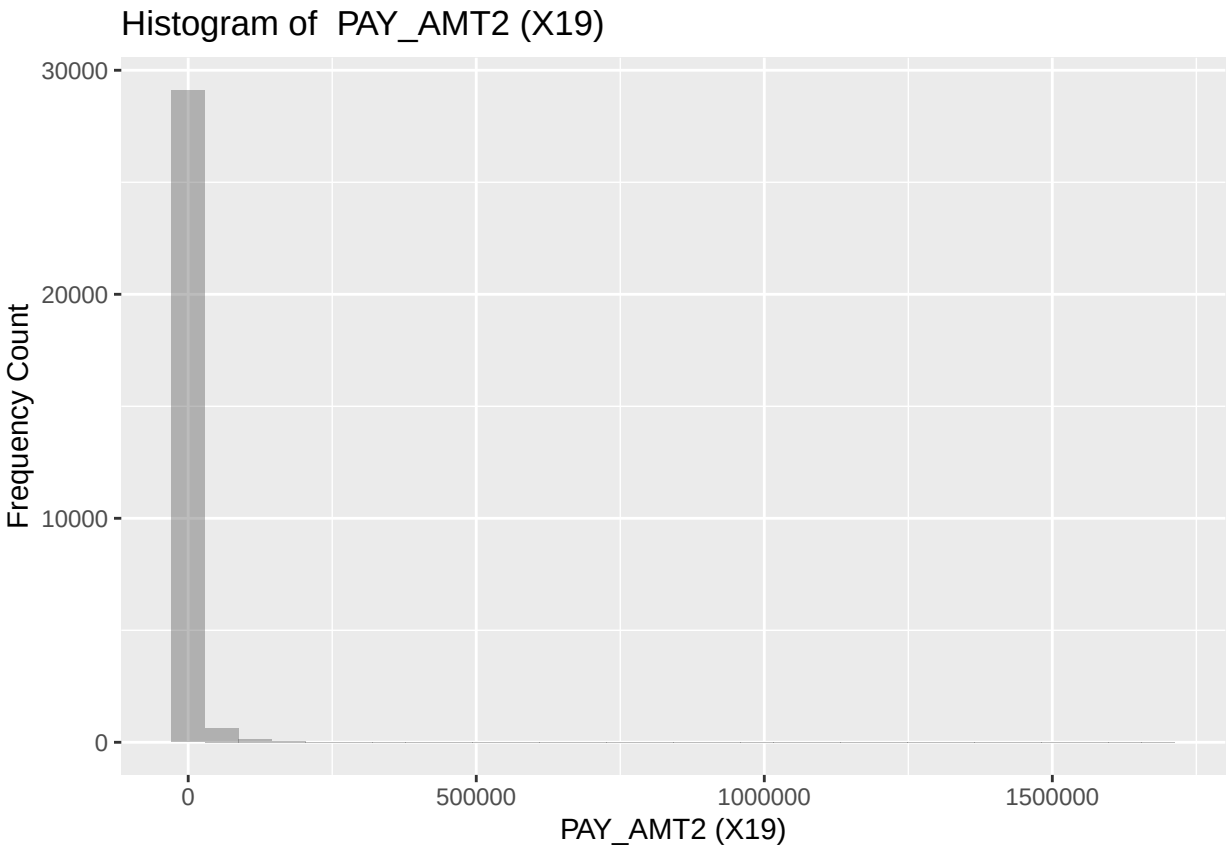
The PAY_AMT2 feature captures the payment amount from the individual for the month of **August, 2005**.

```
summary(df$PAY_AMT2)
```

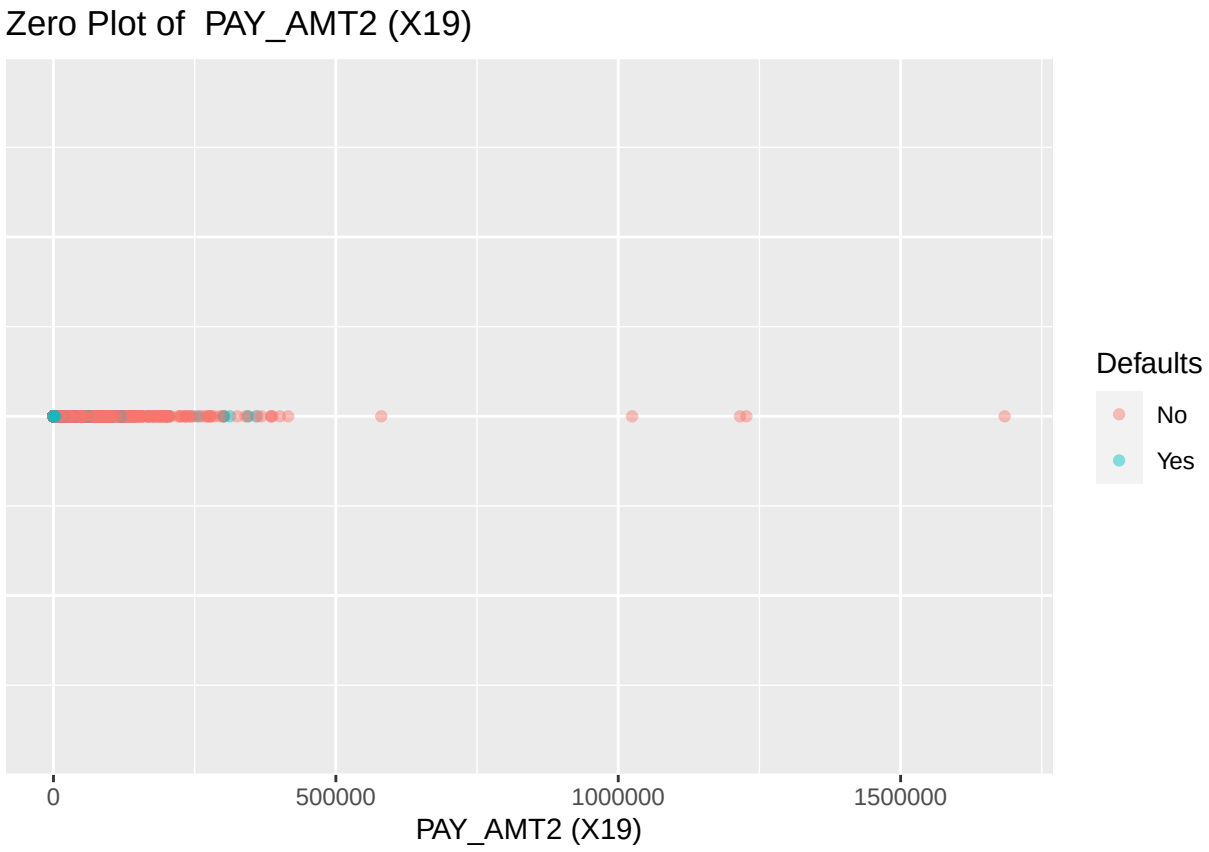
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     833    2009    5921    5000 1684259
```

Graphical Analysis

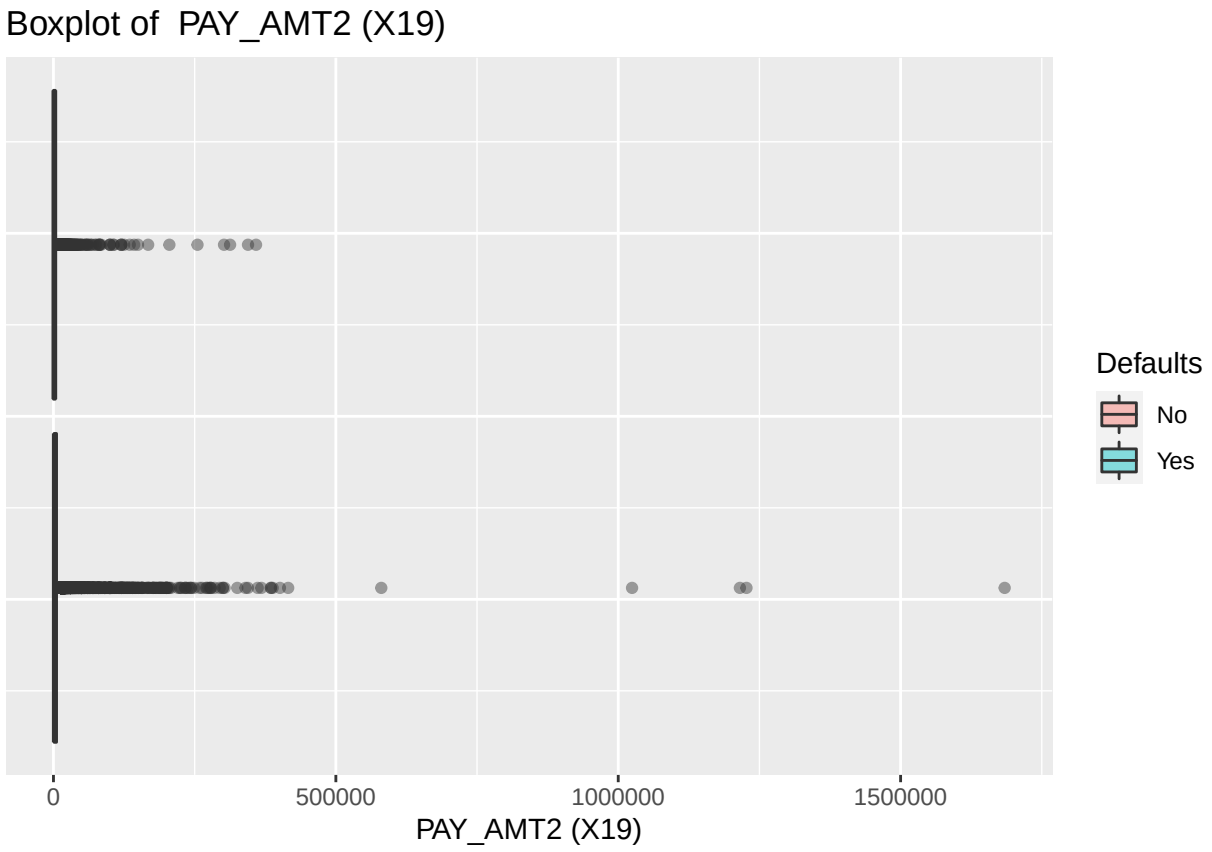
```
x <- df$PAY_AMT2
name <- "PAY_AMT2 (X19)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT3 (X20)

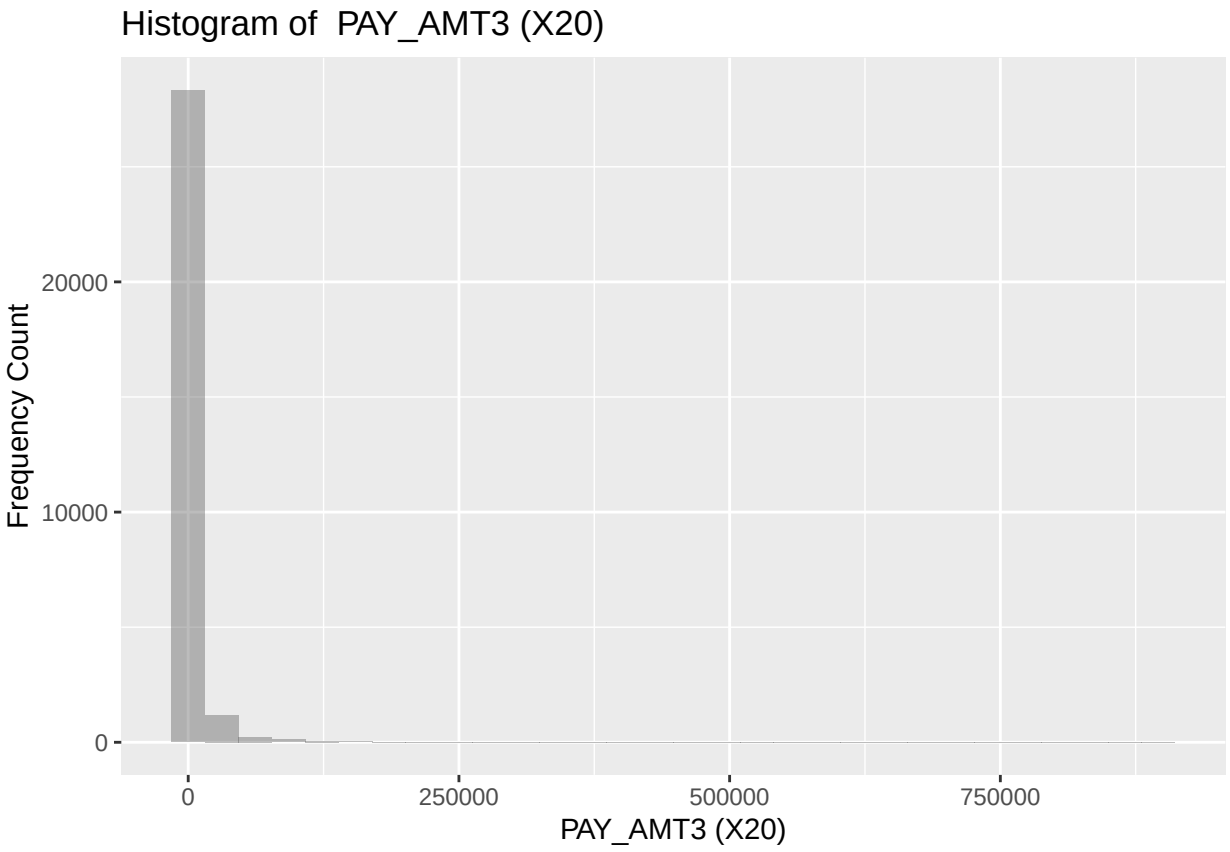
The PAY_AMT3 feature captures the payment amount from the individual for the month of **July, 2005**.

```
summary(df$PAY_AMT3)
```

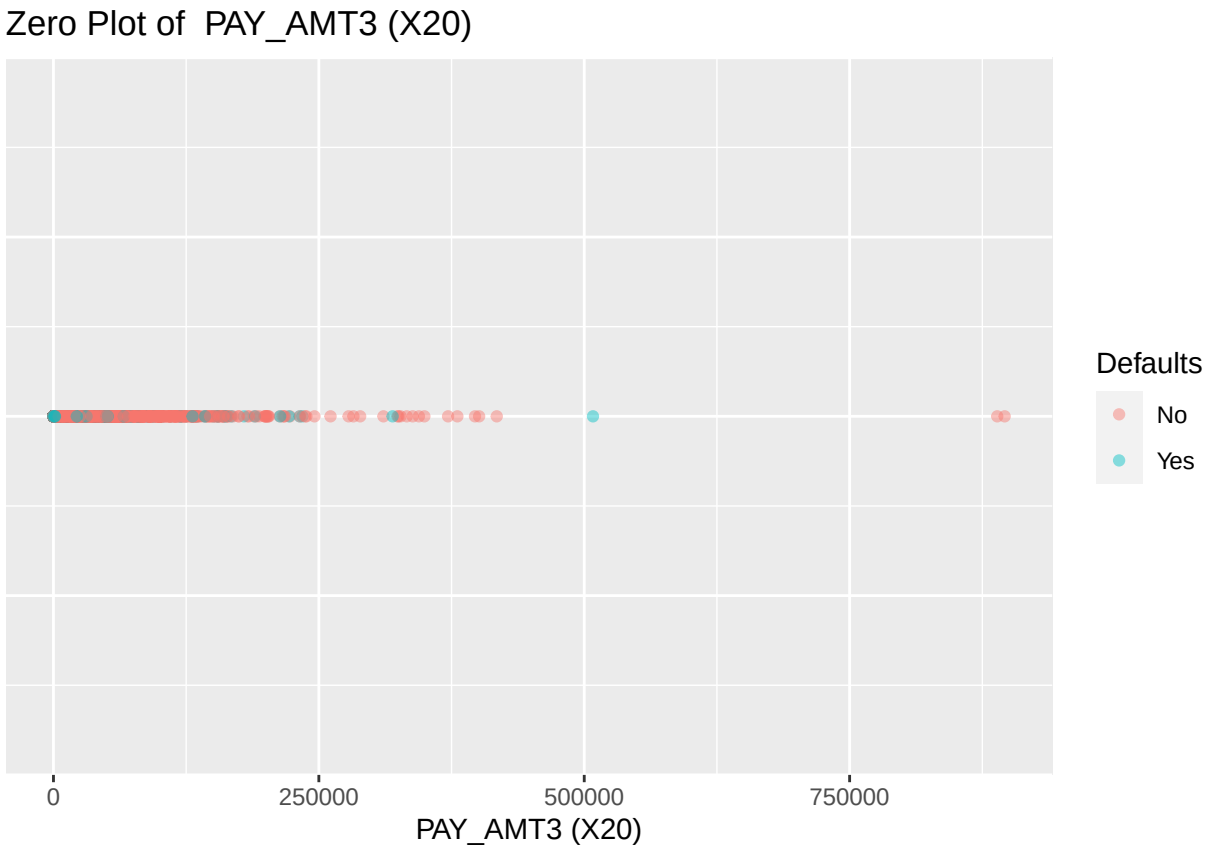
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     390    1800    5226    4505  896040
```

Graphical Analysis

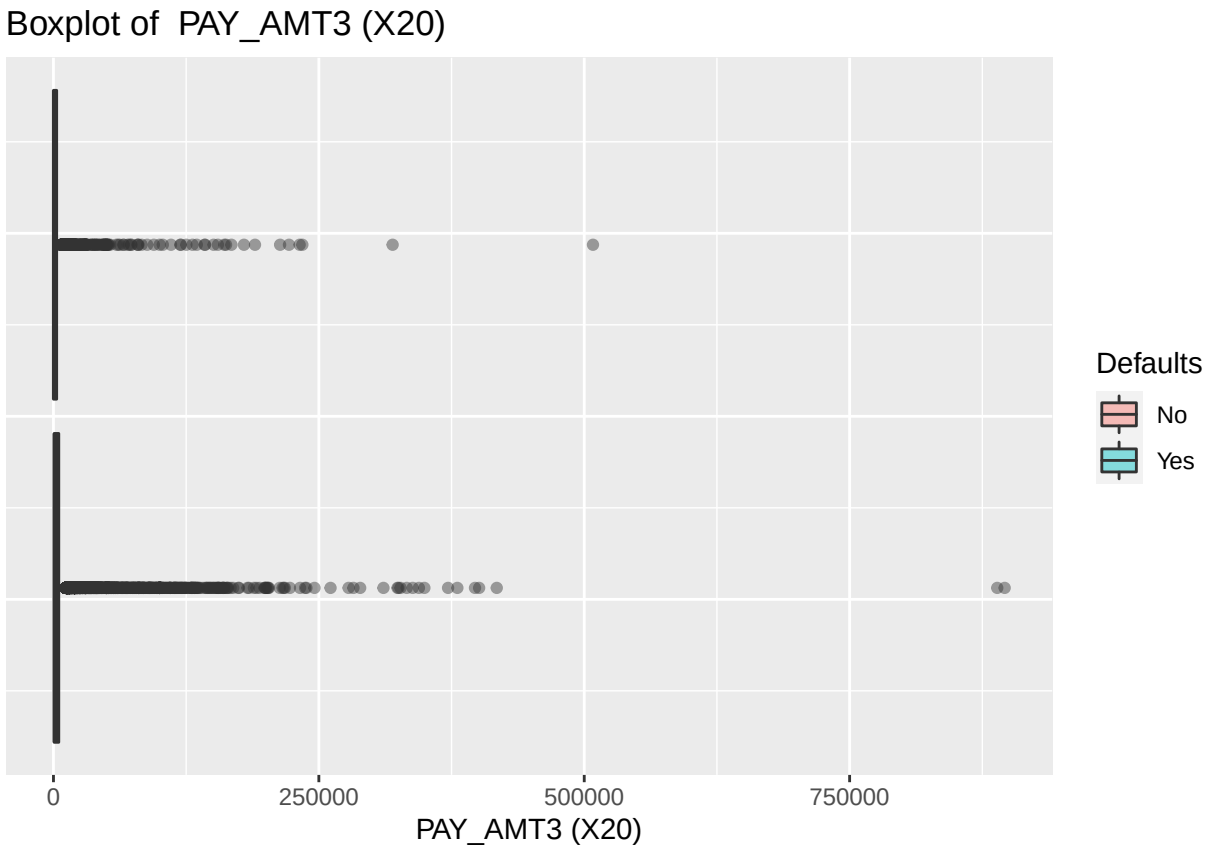
```
x <- df$PAY_AMT3
name <- "PAY_AMT3 (X20)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT4 (X21)

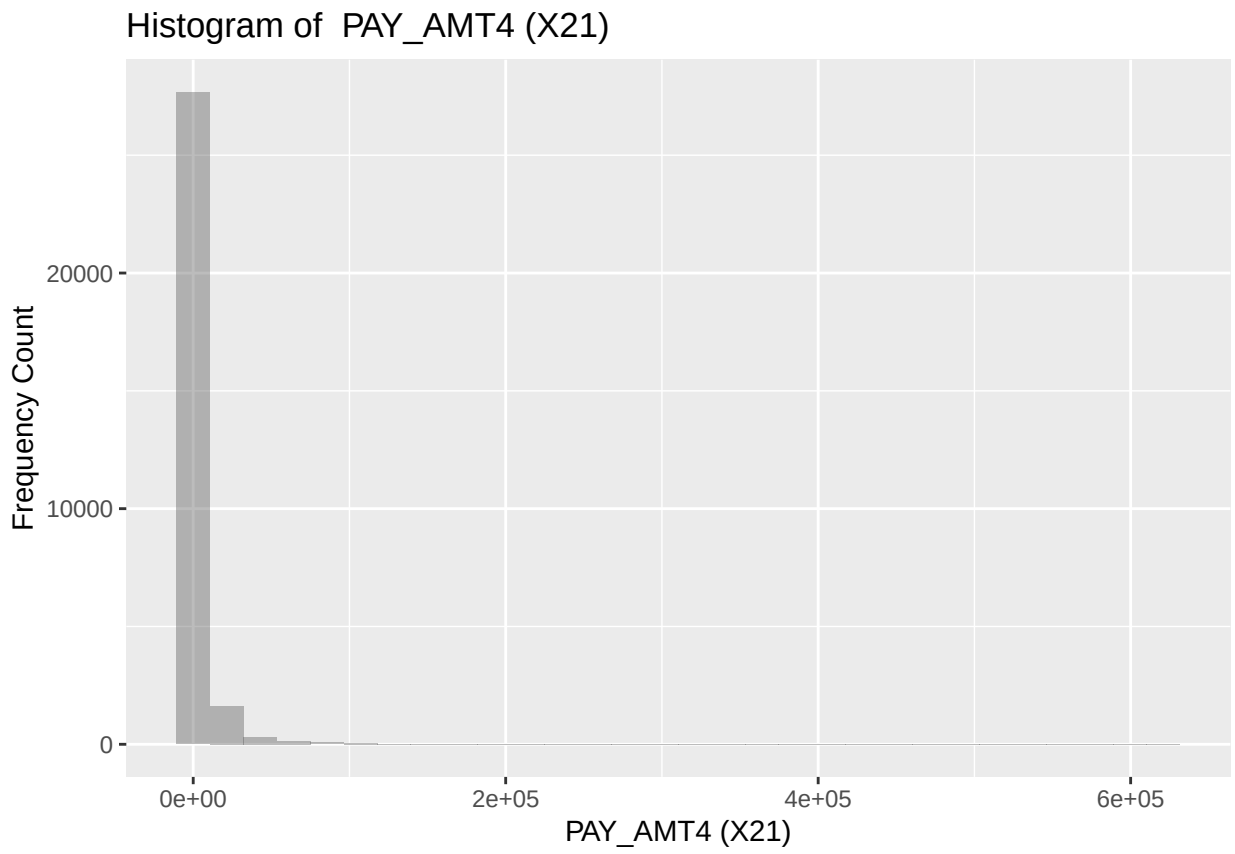
The PAY_AMT4 feature captures the payment amount from the individual for the month of **June, 2005**.

```
summary(df$PAY_AMT4)
```

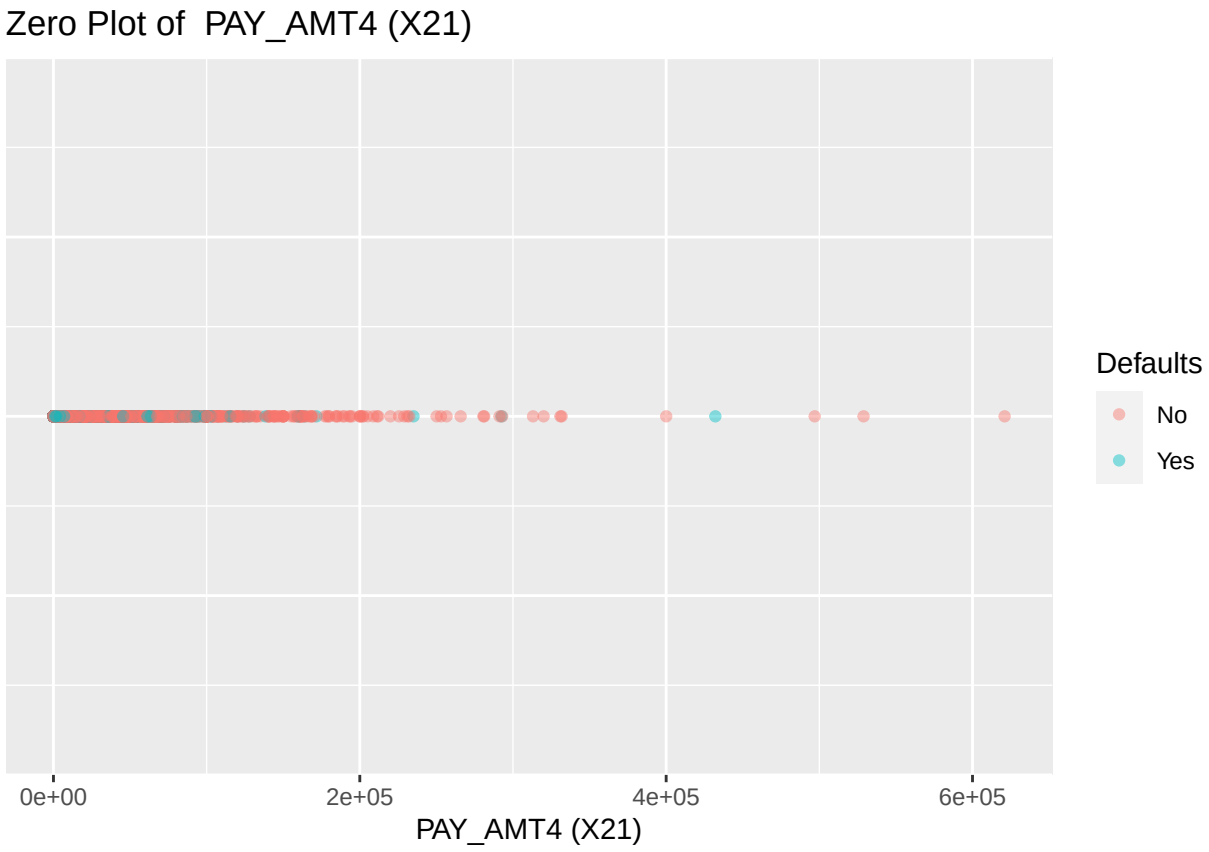
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     296    1500    4826    4013  621000
```

Graphical Analysis

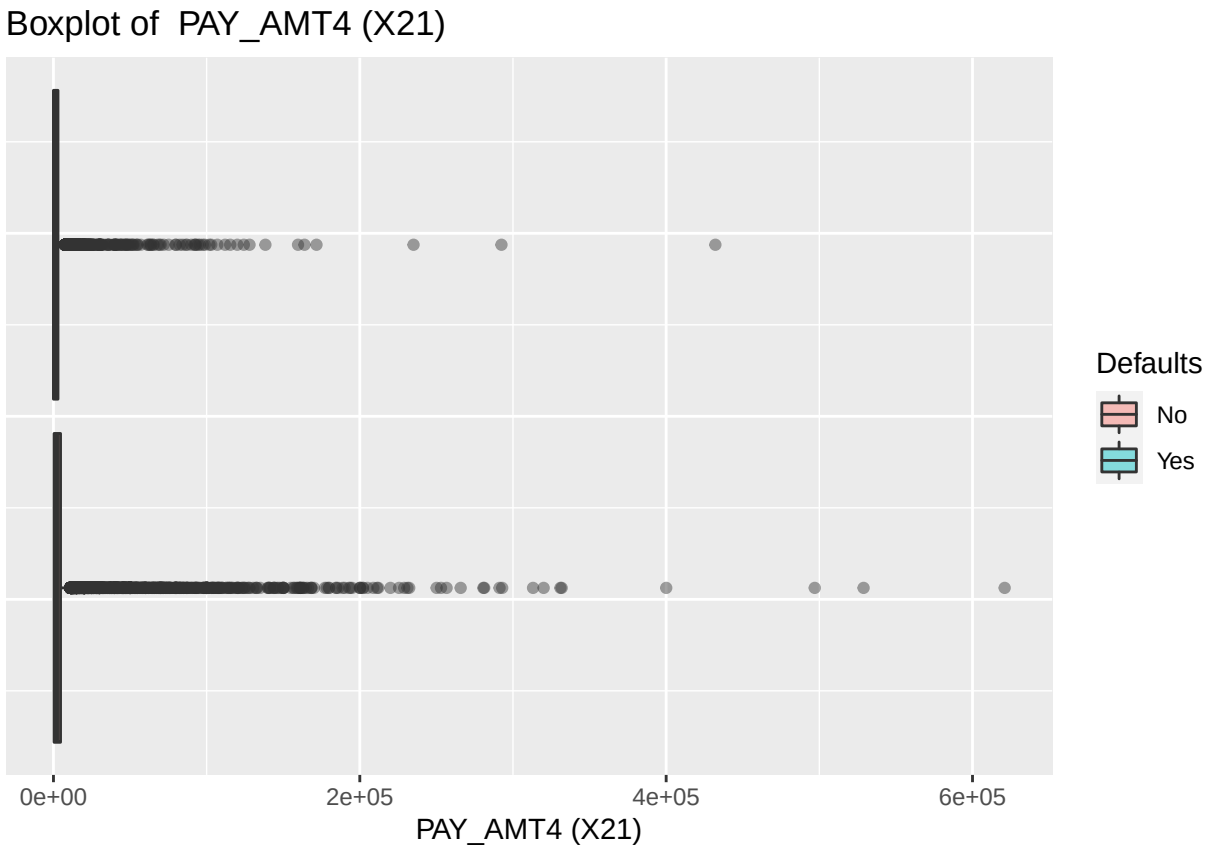
```
x <- df$PAY_AMT4
name <- "PAY_AMT4 (X21)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT5 (X22)

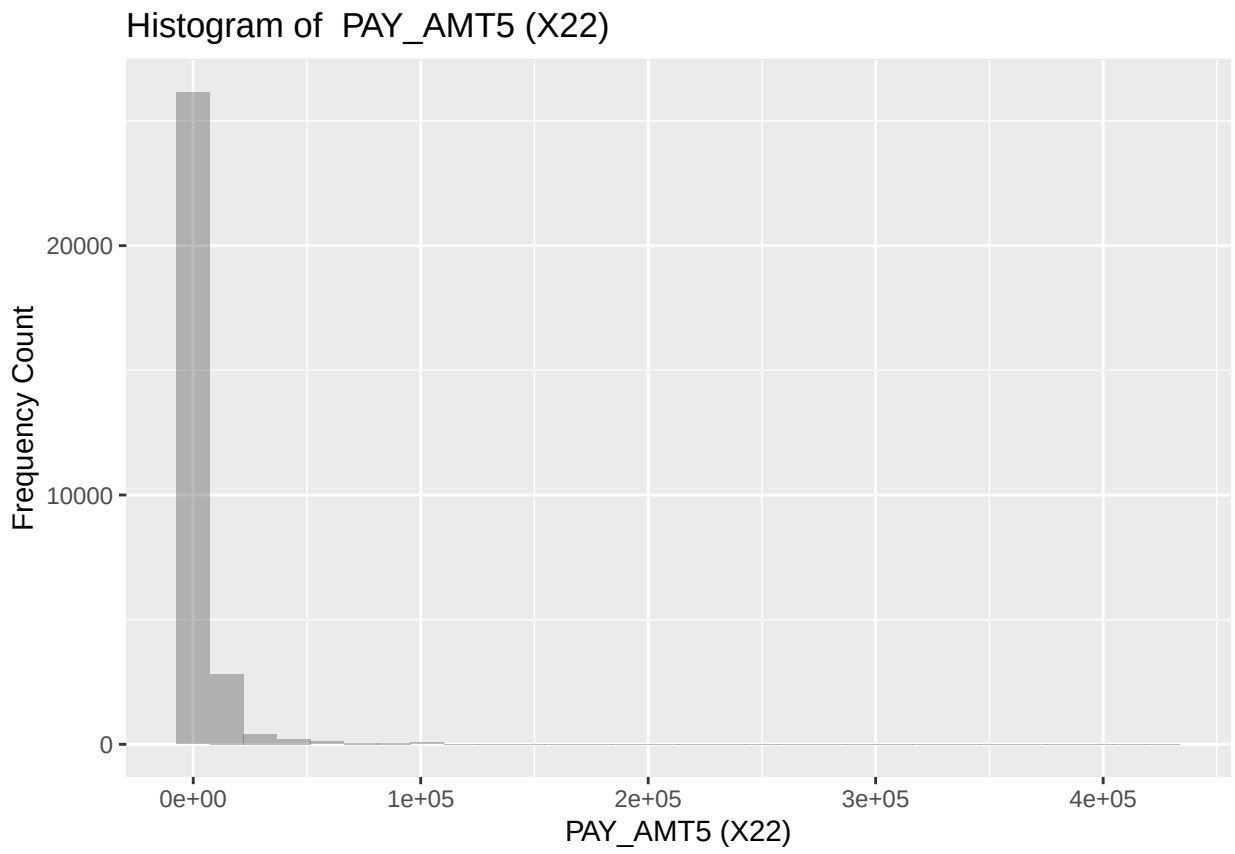
The PAY_AMT5 feature captures the payment amount from the individual for the month of **May, 2005**.

```
summary(df$PAY_AMT5)
```

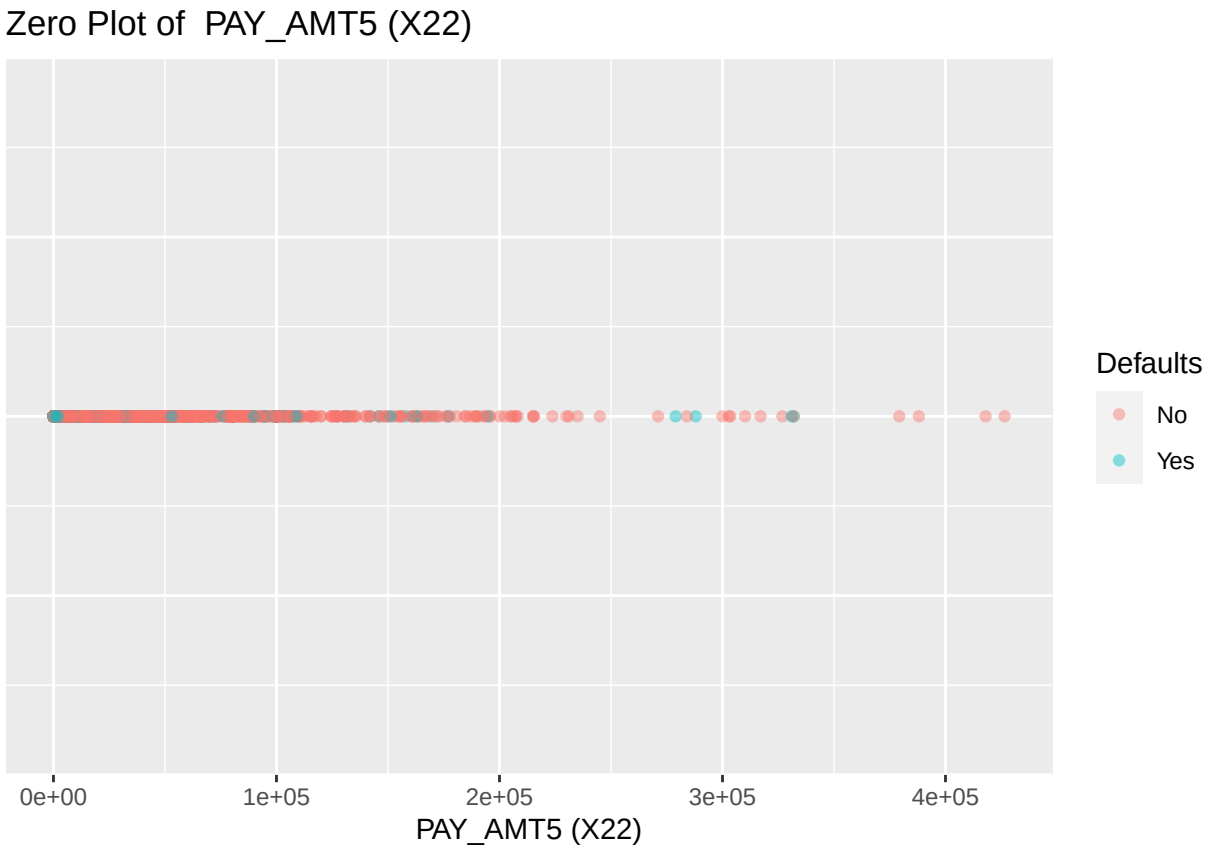
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    252.5   1500.0   4799.4   4031.5 426529.0
```

Graphical Analysis

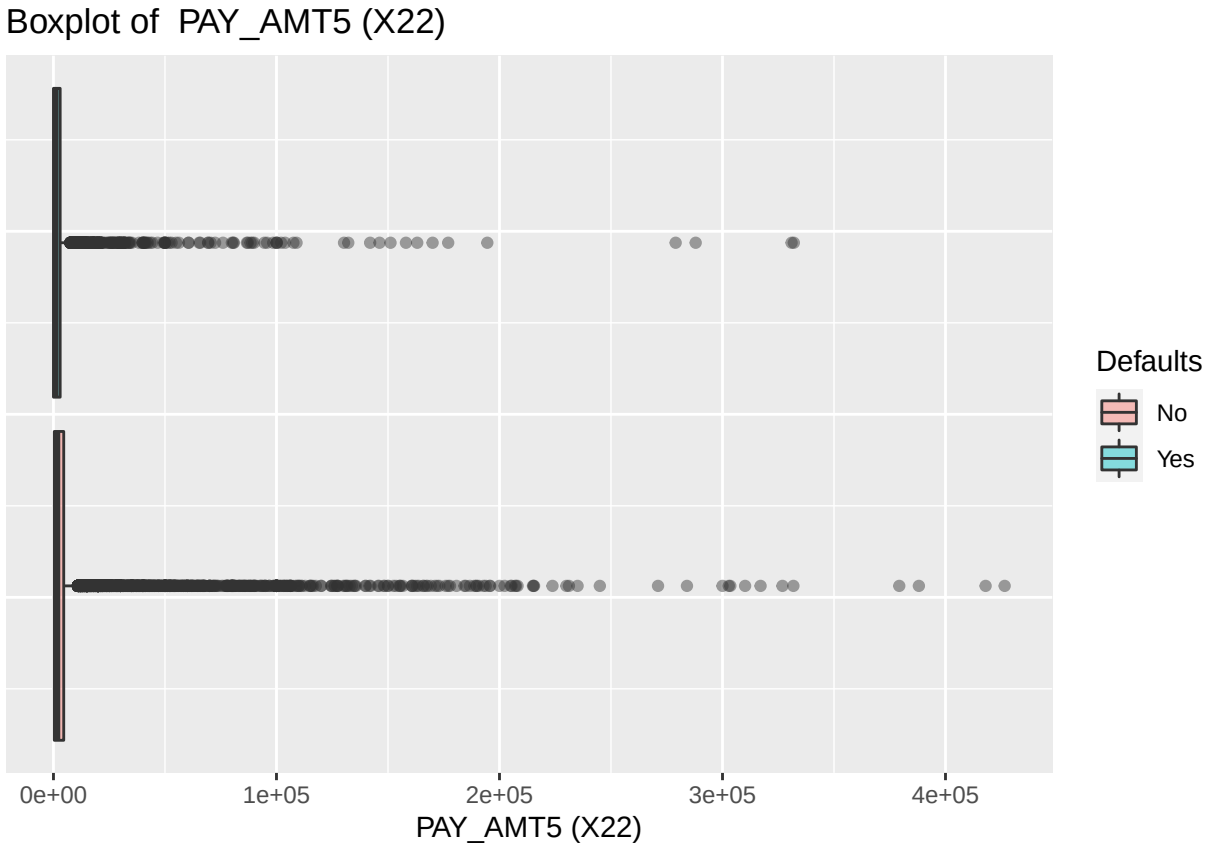
```
x <- df$PAY_AMT5
name <- "PAY_AMT5 (X22)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



PAY_AMT6 (X23)

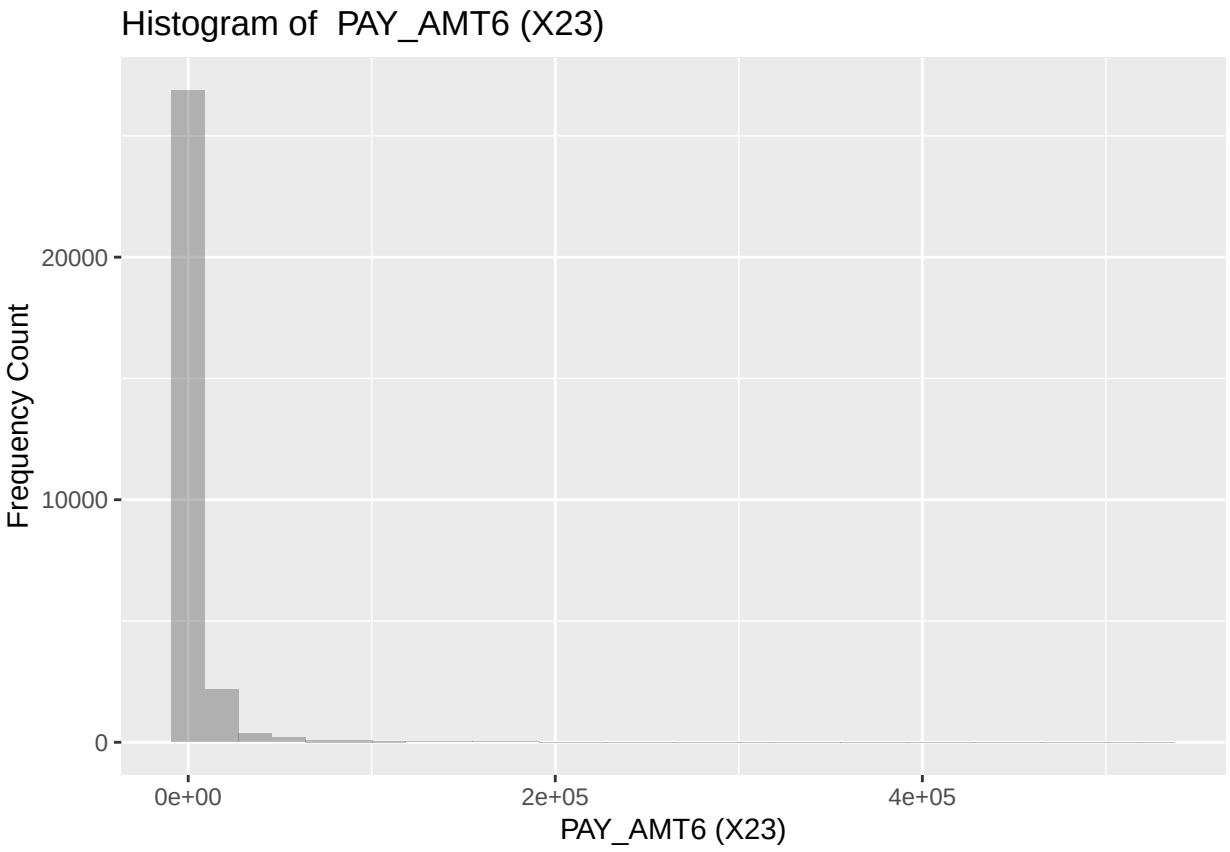
The PAY_AMT6 feature captures the payment amount from the individual for the month of **April, 2005**.

```
summary(df$PAY_AMT6)
```

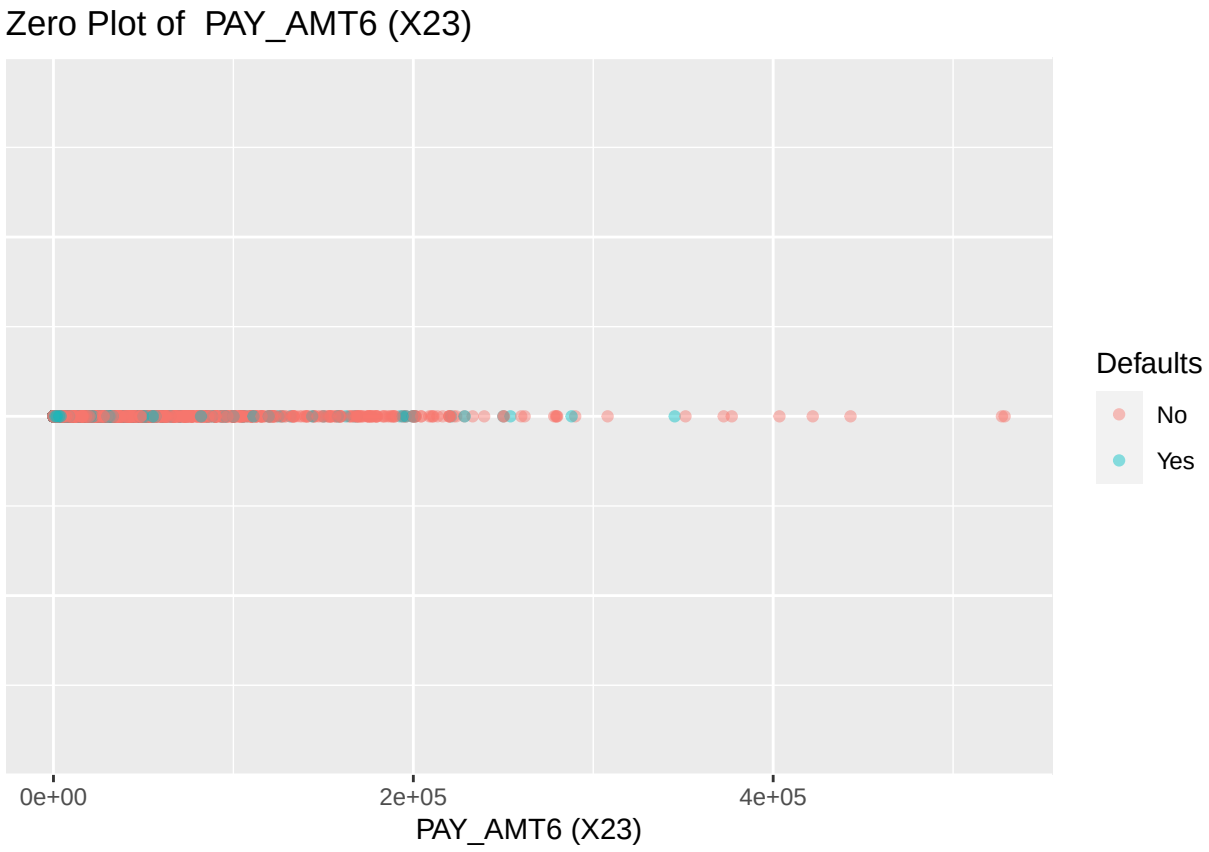
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##      0.0    117.8   1500.0   5215.5   4000.0 528666.0
```

Graphical Analysis

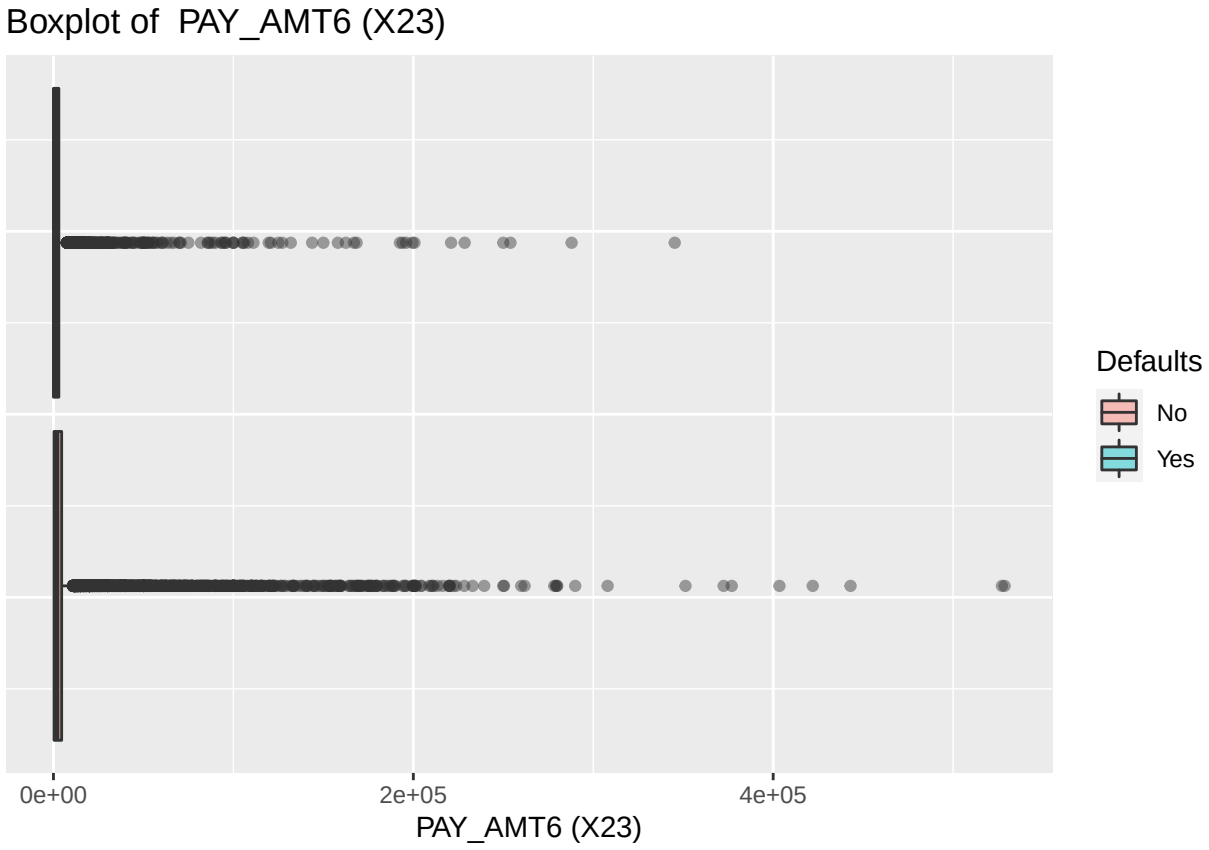
```
x <- df$PAY_AMT6
name <- "PAY_AMT6 (X23)"
base_hist(df,x,name)
```




```
zero_plot(df,x,name)
```




```
boxplot_comp(df,x,name)
```



Defaults (Y1)

The target variable for the dataset is whether the individual defaults on a payment.

```
table(df$default.payment.next.month)
```

```
##
##    No   Yes
## 23364  6636
```