Assignment: 9.2 Final Project Step 2

Raghuwanshi, Prashant

2021-08-07

Introduction

Credit score cards are a common risk control method in the financial industry.

It uses personal information and data submitted by credit card applicants to predict the probability of future defaults and credit card borrowings.

The bank is able to decide whether to issue a credit card to the applicant. Credit scores can objectively quantify the magnitude of risk.

I want to perform below EDA on the top of credit card application data

split out the data and try to found the correlation , dependent variables, independent variables

outliers and biased for the given data set. this will help me researches on the best variables which the credit

card company give more importance while issuing new card to the best fited card member

Research questions

Q find out outliers in all variables and removed outlier records from dataframe

find out the correlation between numeric variables and plot the correlation using various technique

fit liner regression model and find the most significant variables in df

do credit default and applicant annual income have strong significance

do applicant with long employment and age can be consider to qualify for credit card

Approach

How your approach addresses (fully or partially) the problem.

###my approach will focus on below lsited three points ###1 Getting a better understanding of data ###2 Identifying various data patterns ###3 Getting a better understanding of the problem statement ### below steps which requried to address above points ### Checking Introductory Details About Data ### Statistical Insight ### Data cleaning ### Checking Duplicates ### Data Visualization ### Multi-Variate analysisVarious Plots ### Here EDA approach will give the right inofrmation for dependent and multiple independent varibales ### which is going to help me in identifying the closes fit independent variable ### which help researcher to recommend the card types to the cities and gender types

Data (2 Datasets - but no requirement on number of fields or rows)

https://www.kaggle.com/rikdifos/credit-card-approval-prediction?select=credit_record.csv

https://www.kaggle.com/rikdifos/credit-card-approval-prediction?select=application_record.csv

Required Packages

```
library(ggplot2)
theme_set(theme_minimal())
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(QuantPsyc)
```

```
## Loading required package: boot
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'QuantPsyc'
```

```
## The following object is masked from 'package:base':
##
##     norm
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(PerformanceAnalytics)
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Plots and Table Needs

###Correlation Matrix ###Scatterplot ###pairplot ###Histogram ###Boxplot

Questions for future steps

find out the correlation among variables

club application and credit report for current month

perform EDA by using R programming

How to import and clean my data

```
### Set the working directory to the root of your DSC 520 directory
setwd("D:/MS_DataScience/DSC 520-Datastatiics.pdf/dsc520")
### Load card data `data/credit_record.csv` to DF
credit_df <- read.csv("data/credit_record.csv")
### Cleaning records, by filtering out old records and keep current credit record
credit_val <- filter(credit_df, MONTHS_BALANCE==0)
head(credit_val)
```

```
##        ID MONTHS_BALANCE STATUS
## 1 5001711              0      X
## 2 5001712              0      C
## 3 5001713              0      X
## 4 5001714              0      X
## 5 5001715              0      X
## 6 5001717              0      C
```

```
### Load credit card application data `data/application_record.csv` to DF
appl_df <- read.csv("data/application_record.csv")

### Join the credit data with application data and find out current credit status
card_df <- merge(x = appl_df, y = credit_val, by = "ID", all.x = TRUE)
head(card_df)
```

```
##        ID CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY CNT_CHILDREN
## 1 5008804           M            Y               Y            0
## 2 5008805           M            Y               Y            0
## 3 5008806           M            Y               Y            0
## 4 5008808           F            N               Y            0
## 5 5008809           F            N               Y            0
## 6 5008810           F            N               Y            0
##   AMT_INCOME_TOTAL     NAME_INCOME_TYPE           NAME_EDUCATION_TYPE
## 1           427500              Working              Higher education
## 2           427500              Working              Higher education
## 3           112500              Working Secondary / secondary special
## 4           270000 Commercial associate Secondary / secondary special
## 5           270000 Commercial associate Secondary / secondary special
## 6           270000 Commercial associate Secondary / secondary special
##     NAME_FAMILY_STATUS NAME_HOUSING_TYPE DAYS_BIRTH DAYS_EMPLOYED FLAG_MOBIL
## 1       Civil marriage  Rented apartment     -12005         -4542          1
## 2       Civil marriage  Rented apartment     -12005         -4542          1
## 3              Married House / apartment     -21474         -1134          1
## 4 Single / not married House / apartment     -19110         -3051          1
## 5 Single / not married House / apartment     -19110         -3051          1
## 6 Single / not married House / apartment     -19110         -3051          1
##   FLAG_WORK_PHONE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS
## 1               1          0          0                              2
```

```
## 2               1          0          0                             2
## 3               0          0          0  Security staff             2
## 4               0          1          1     Sales staff             1
## 5               0          1          1     Sales staff             1
## 6               0          1          1     Sales staff             1
##   MONTHS_BALANCE STATUS
## 1              0      C
## 2              0      C
## 3              0      C
## 4              0      0
## 5             NA   <NA>
## 6              0      C
```

```
### fixing missing value,defaulting the missed month balance & status as 0
card_df$MONTHS_BALANCE <- factor(ifelse( is.na(card_df$MONTHS_BALANCE), 0, card_df$MONTHS_BALANCE))
card_df$STATUS <- factor(ifelse( is.na(card_df$STATUS), 0, card_df$STATUS))
### factoring variables
card_df$CODE_GENDER <- factor(card_df$CODE_GENDER)
card_df$FLAG_OWN_CAR <- factor(card_df$FLAG_OWN_CAR)
card_df$FLAG_OWN_REALTY <- factor(card_df$FLAG_OWN_REALTY)
card_df$NAME_EDUCATION_TYPE <- factor(card_df$NAME_EDUCATION_TYPE)
```

What does the final data set look like?

Preparing new data set with required City, Date, Card.Type, Exp.Type, Gender, Amount, Tier fields

```
head(card_df)
```

```
##        ID CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY CNT_CHILDREN
## 1 5008804           M            Y               Y            0
## 2 5008805           M            Y               Y            0
## 3 5008806           M            Y               Y            0
## 4 5008808           F            N               Y            0
## 5 5008809           F            N               Y            0
## 6 5008810           F            N               Y            0
##   AMT_INCOME_TOTAL     NAME_INCOME_TYPE           NAME_EDUCATION_TYPE
## 1           427500              Working              Higher education
## 2           427500              Working              Higher education
## 3           112500              Working Secondary / secondary special
## 4           270000 Commercial associate Secondary / secondary special
## 5           270000 Commercial associate Secondary / secondary special
## 6           270000 Commercial associate Secondary / secondary special
##     NAME_FAMILY_STATUS NAME_HOUSING_TYPE DAYS_BIRTH DAYS_EMPLOYED FLAG_MOBIL
## 1       Civil marriage  Rented apartment     -12005         -4542          1
## 2       Civil marriage  Rented apartment     -12005         -4542          1
## 3              Married House / apartment     -21474         -1134          1
## 4 Single / not married House / apartment     -19110         -3051          1
## 5 Single / not married House / apartment     -19110         -3051          1
## 6 Single / not married House / apartment     -19110         -3051          1
##   FLAG_WORK_PHONE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS
## 1               1          0          0                             2
## 2               1          0          0                             2
## 3               0          0          0  Security staff             2
```

```
## 4               0          1          1     Sales staff               1
## 5               0          1          1     Sales staff               1
## 6               0          1          1     Sales staff               1
##   MONTHS_BALANCE STATUS
## 1              0      C
## 2              0      C
## 3              0      C
## 4              0      0
## 5              0      0
## 6              0      C
```

```
str(card_df)
```

```
## 'data.frame':    438557 obs. of  20 variables:
##  $ ID                 : int  5008804 5008805 5008806 5008808 5008809 5008810 5008811 5008812 5008813
##  $ CODE_GENDER        : Factor w/ 2 levels "F","M": 2 2 2 1 1 1 1 1 1 1 ...
##  $ FLAG_OWN_CAR       : Factor w/ 2 levels "N","Y": 2 2 2 1 1 1 1 1 1 1 ...
##  $ FLAG_OWN_REALTY    : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ CNT_CHILDREN       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_INCOME_TOTAL   : num  427500 427500 112500 270000 270000 ...
##  $ NAME_INCOME_TYPE   : chr  "Working" "Working" "Working" "Commercial associate" ...
##  $ NAME_EDUCATION_TYPE: Factor w/ 5 levels "Academic degree",..: 2 2 5 5 5 5 5 2 2 2 ...
##  $ NAME_FAMILY_STATUS : chr  "Civil marriage" "Civil marriage" "Married" "Single / not married" ...
##  $ NAME_HOUSING_TYPE  : chr  "Rented apartment" "Rented apartment" "House / apartment" "House / apar
##  $ DAYS_BIRTH         : int  -12005 -12005 -21474 -19110 -19110 -19110 -19110 -22464 -22464 -22464 .
##  $ DAYS_EMPLOYED      : int  -4542 -4542 -1134 -3051 -3051 -3051 -3051 365243 365243 365243 ...
##  $ FLAG_MOBIL         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_WORK_PHONE    : int  1 1 0 0 0 0 0 0 0 0 ...
##  $ FLAG_PHONE         : int  0 0 0 1 1 1 1 0 0 0 ...
##  $ FLAG_EMAIL         : int  0 0 0 1 1 1 1 0 0 0 ...
##  $ OCCUPATION_TYPE    : chr  "" "" "Security staff" "Sales staff" ...
##  $ CNT_FAM_MEMBERS    : num  2 2 2 1 1 1 1 1 1 1 ...
##  $ MONTHS_BALANCE     : Factor w/ 1 level "0": 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATUS             : Factor w/ 8 levels "0","1","2","3",..: 7 7 7 1 1 7 7 1 1 1 ...
```

```
dim(card_df)
```

```
## [1] 438557     20
```

```
summary(card_df)
```

```
##        ID          CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY  CNT_CHILDREN
##  Min.   :5008804   F:294440    N:275459     N:134483        Min.   : 0.0000
##  1st Qu.:5609375   M:144117    Y:163098     Y:304074        1st Qu.: 0.0000
##  Median :6047745                                          Median : 0.0000
##  Mean   :6022176                                          Mean   : 0.4274
##  3rd Qu.:6456971                                          3rd Qu.: 1.0000
##  Max.   :7999952                                          Max.   :19.0000
##
##  AMT_INCOME_TOTAL  NAME_INCOME_TYPE                     NAME_EDUCATION_TYPE
##  Min.   :  26100   Length:438557      Academic degree              :   312
##  1st Qu.: 121500   Class :character   Higher education             :117522
##  Median : 160780   Mode  :character   Incomplete higher            : 14851
##  Mean   : 187524                      Lower secondary              :  4051
##  3rd Qu.: 225000                      Secondary / secondary special:301821
##  Max.   :6750000
```

```
##
##  NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH     DAYS_EMPLOYED
##  Length:438557      Length:438557      Min.   :-25201   Min.   :-17531
##  Class :character   Class :character   1st Qu.:-19483   1st Qu.: -3103
##  Mode  :character   Mode  :character   Median :-15630   Median : -1467
##                                        Mean   :-15998   Mean   : 60564
##                                        3rd Qu.:-12514   3rd Qu.:  -371
##                                        Max.   : -7489   Max.   :365243
##
##    FLAG_MOBIL FLAG_WORK_PHONE    FLAG_PHONE       FLAG_EMAIL
##  Min.   :1    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1    Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :1    Mean   :0.2061   Mean   :0.2878   Mean   :0.1082
##  3rd Qu.:1    3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS  MONTHS_BALANCE     STATUS
##  Length:438557      Min.   : 1.000   0:438557       0      :420771
##  Class :character   1st Qu.: 2.000                  C      : 12974
##  Mode  :character   Median : 2.000                  X      :  4487
##                     Mean   : 2.194                  1      :   236
##                     3rd Qu.: 3.000                  5      :    59
##                     Max.   :20.000                  2      :    19
##                                                     (Other):    11
```

How do you plan to slice and dice the data?

```
card_df1 <- dplyr::select(card_df, ID, CNT_CHILDREN, AMT_INCOME_TOTAL, DAYS_BIRTH, DAYS_EMPLOYED, STATUS
card_df2 <- filter(card_df1, STATUS==1)
card_df3 <- dplyr::select(card_df2, CNT_CHILDREN, AMT_INCOME_TOTAL, DAYS_BIRTH, DAYS_EMPLOYED)
dim(card_df3)
```

```
## [1] 236   4
```

Questions for future steps

Q find out outliers in all variables and removed outlier records from dataframe

find out the correlation between numeric variables and plot the correlation using various technique

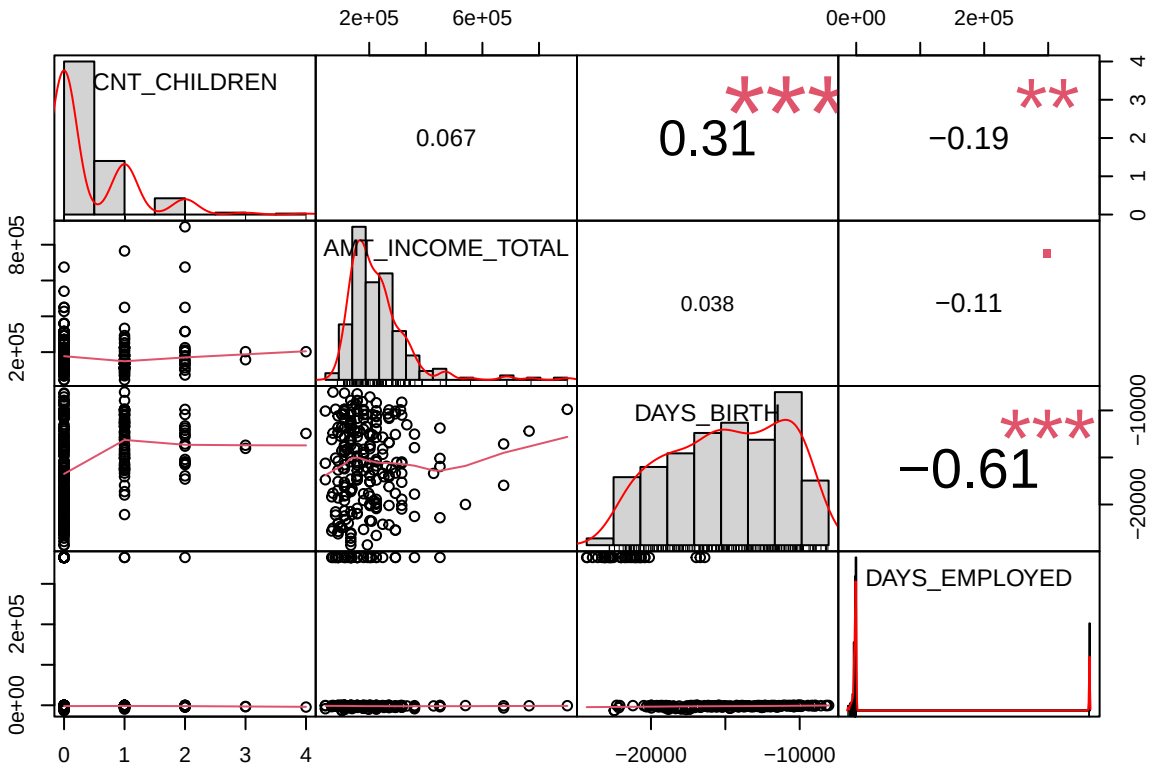
fit liner regression model and find the most significant variables in df

do credit default and applicant annual income have strong significance

do applicant with long employment and age can be consider to qualify for credit card

What types of plots and tables will help you to illustrate the findings to your questions?

```
### Correlation Matrix
chart.Correlation(card_df3, histogram=TRUE, pch=19)
```




```
### In the above plot:
### The distribution of each variable is shown on the diagonal.
### On the bottom of the diagonal : the bivariate scatter plots with a fitted line are displayed
### On the top of the diagonal : the value of the correlation plus the significance level as stars
### Each significance level is associated to a symbol : p-values(0, 0.001, 0.01, 0.05, 0.1, 1) <=> symb
```

correlation

```
res <- cor(card_df3)
round(res, 2)
```

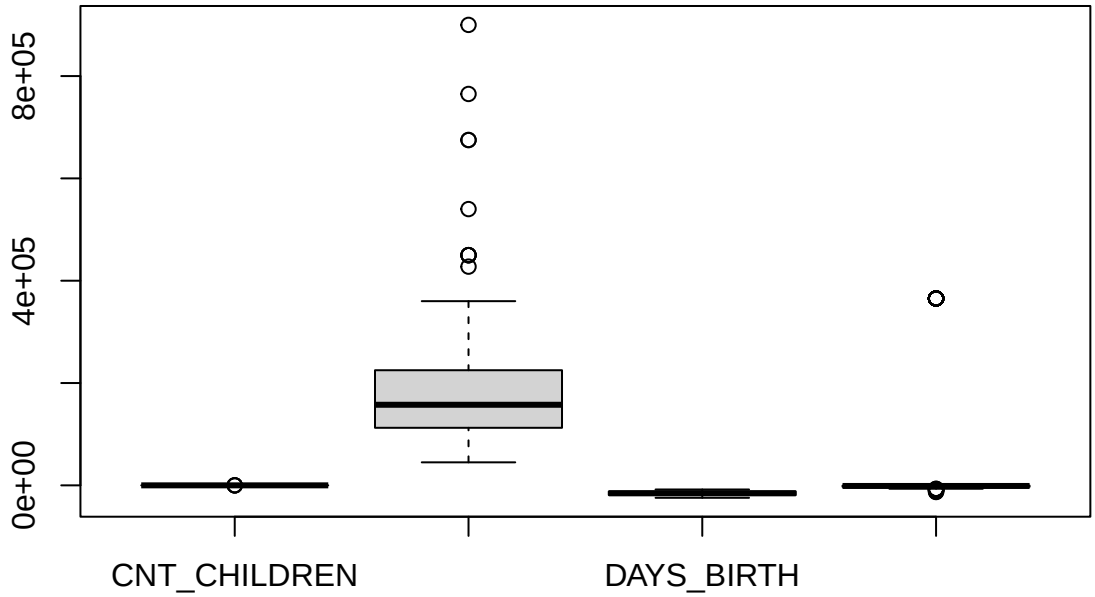
```
##                  CNT_CHILDREN AMT_INCOME_TOTAL DAYS_BIRTH DAYS_EMPLOYED
## CNT_CHILDREN             1.00              0.07       0.31         -0.19
## AMT_INCOME_TOTAL         0.07              1.00       0.04         -0.11
## DAYS_BIRTH               0.31              0.04       1.00         -0.61
## DAYS_EMPLOYED           -0.19             -0.11      -0.61          1.00
```

corrplot

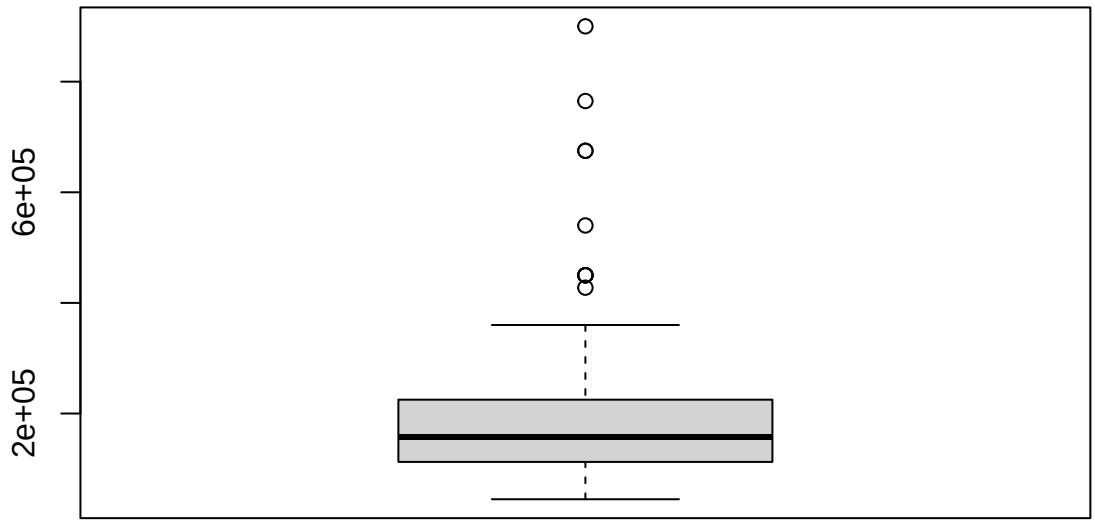
```
library(corrplot)
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)

### Positive correlations are displayed in blue and negative correlations in red color.
### Color intensity and the size of the circle are proportional to the correlation coefficients.
### In the right side of the correlogram, the legend color shows the correlation coefficients and the c
```

boxplot to find out outlier and removed outlier value from my dataset

```
boxplot(card_df3)
```




```
boxplot(card_df3$AMT_INCOME_TOTAL)
```



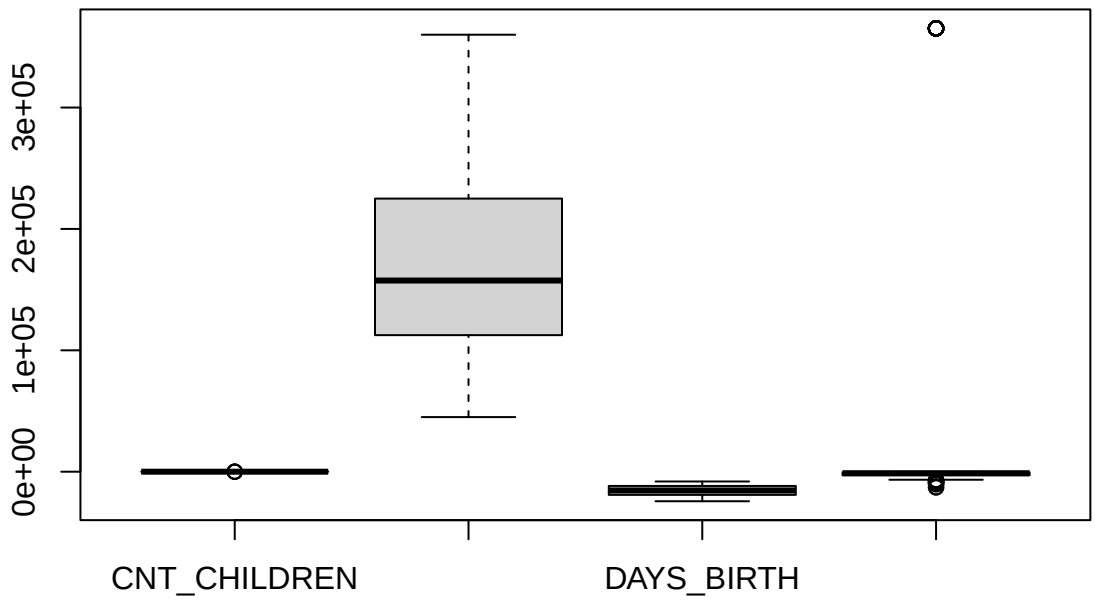

```
### Now you can assign the outlier values into a vector
outliers_AMT_INCOME_TOTAL <- boxplot(card_df3$AMT_INCOME_TOTAL, plot=FALSE)$out
### Check the results
print(outliers_AMT_INCOME_TOTAL)
```

```
##  [1] 765000 540000 450000 675000 675000 900000 450000 427500 450000 450000
```

```
### First you need find in which rows the outliers are
card_df3[which(card_df3$AMT_INCOME_TOTAL %in% outliers_AMT_INCOME_TOTAL),]
```

```
##     CNT_CHILDREN AMT_INCOME_TOTAL DAYS_BIRTH DAYS_EMPLOYED
## 2              1           765000     -12197         -1194
## 34             0           540000     -19996          -691
## 58             2           450000     -11870          -221
## 97             0           675000     -17964         -6594
## 99             2           675000     -13567         -1175
## 112            2           900000      -9889         -1000
## 164            1           450000     -15098         -1562
## 193            0           427500     -16691         -1565
## 203            0           450000     -21406        365243
## 215            0           450000     -15960         -3574
```

```
card_df4 <- card_df3[-which(card_df3$AMT_INCOME_TOTAL %in% outliers_AMT_INCOME_TOTAL),]
### If you check now with boxplot, you will notice that those pesky outliers are gone
boxplot(card_df4)
```



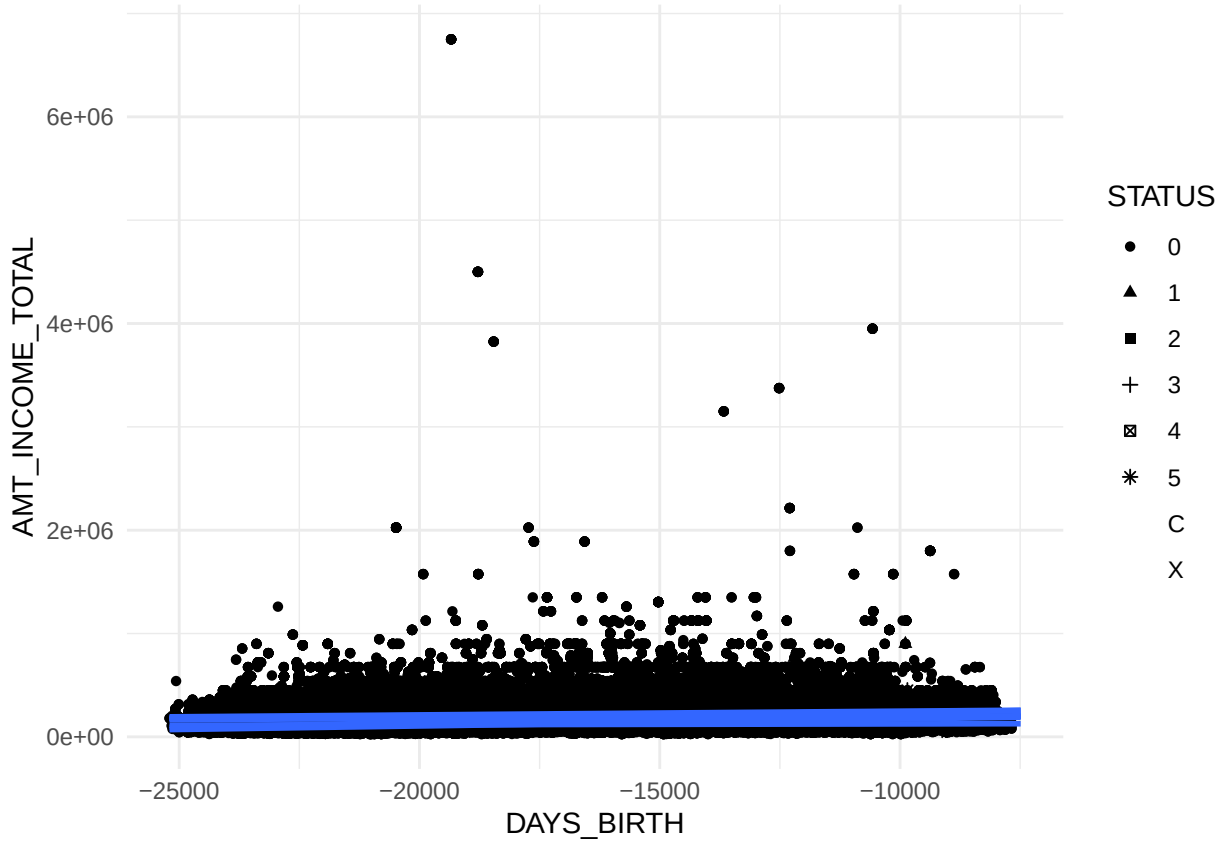
Scatter plots with multiple groups

```
# Change point shapes by the levels of STATUS
ggplot(card_df1, aes(x=DAYS_BIRTH, y=AMT_INCOME_TOTAL, shape=STATUS)) +
  geom_point() + geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

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

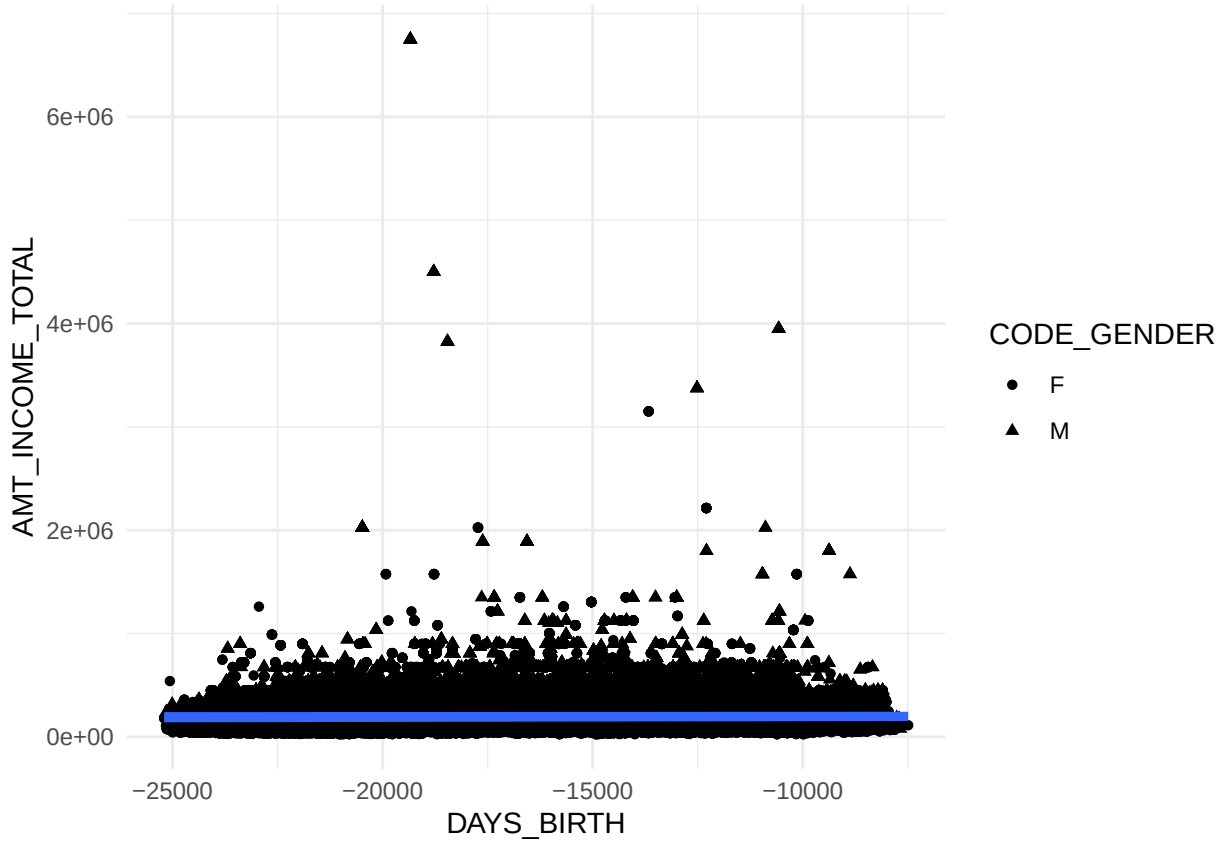
```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 8. Consider
## specifying shapes manually if you must have them.
```

```
## Warning: Removed 17461 rows containing missing values (geom_point).
```



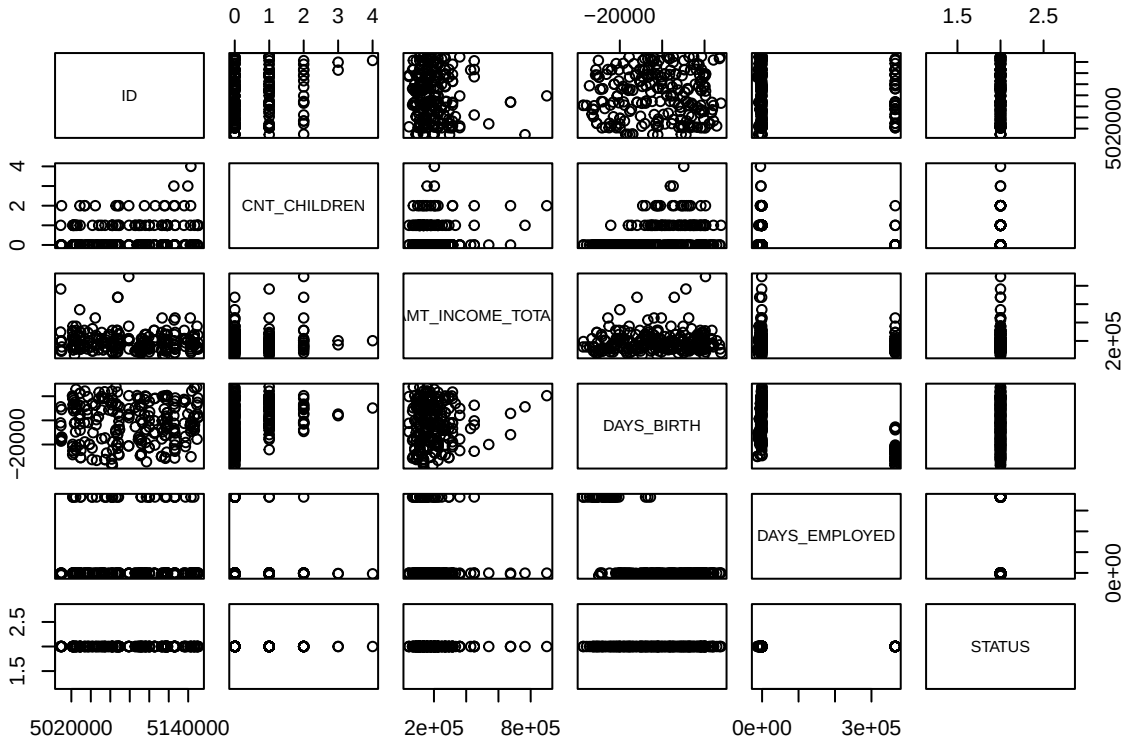

```
# Change point shapes by the levels of STATUS
ggplot(card_df, aes(x=DAYS_BIRTH, y=AMT_INCOME_TOTAL, shape=CODE_GENDER)) +
  geom_point() + geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

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



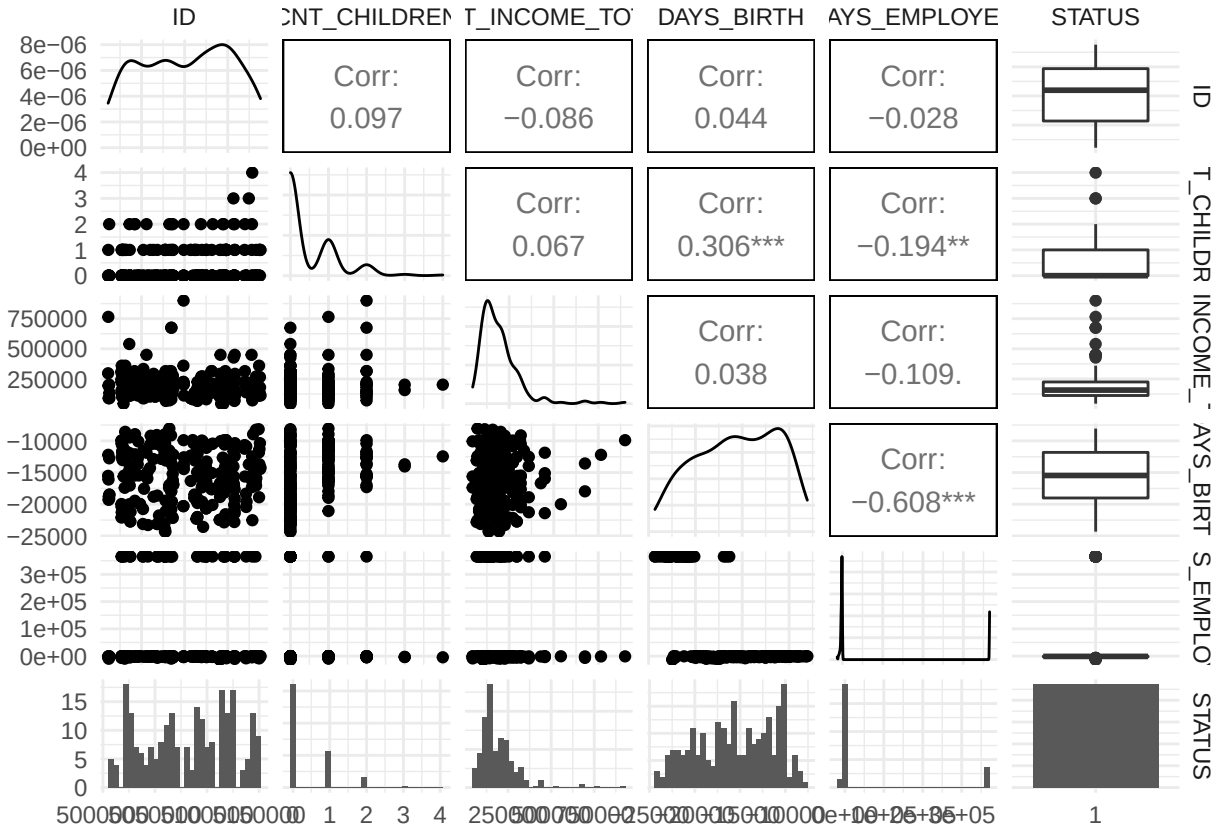
Pairplot

```
pairs(card_df2)
```

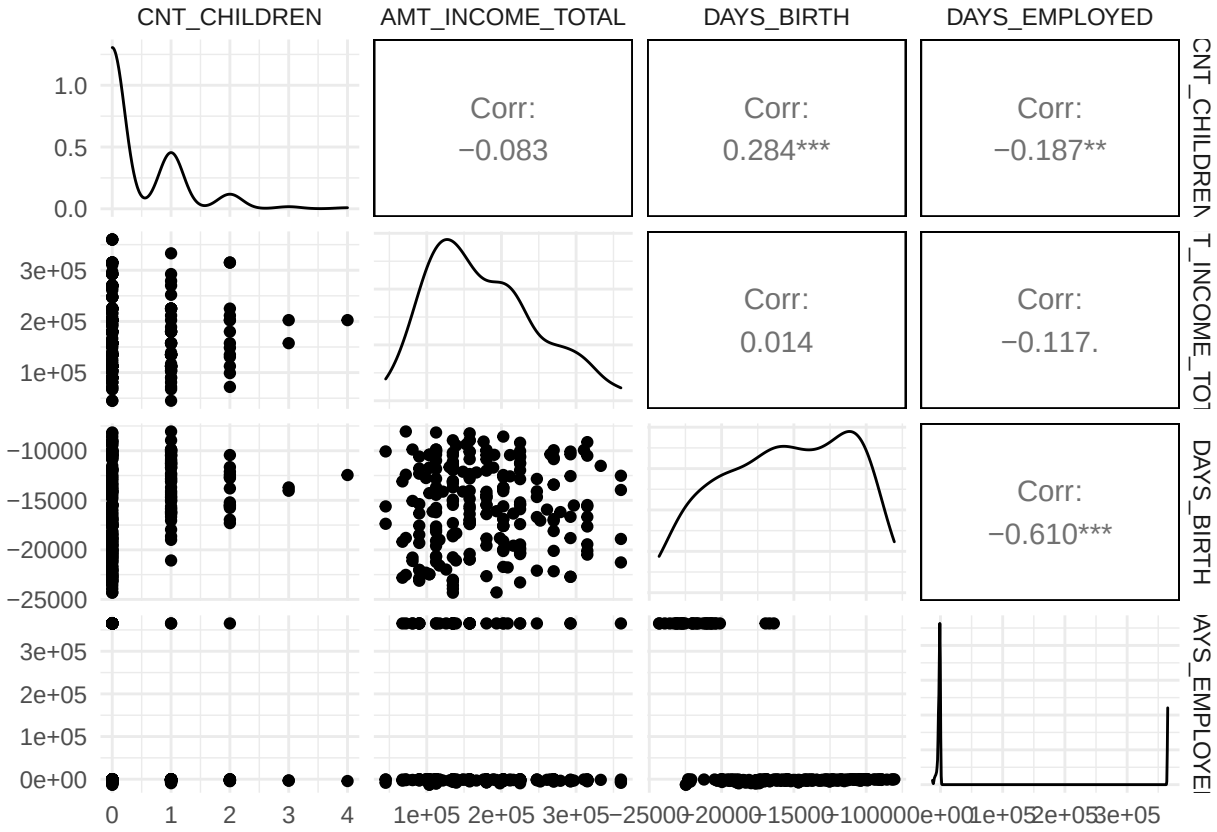



```
ggpairs(card_df2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggpairs(card_df4)
```



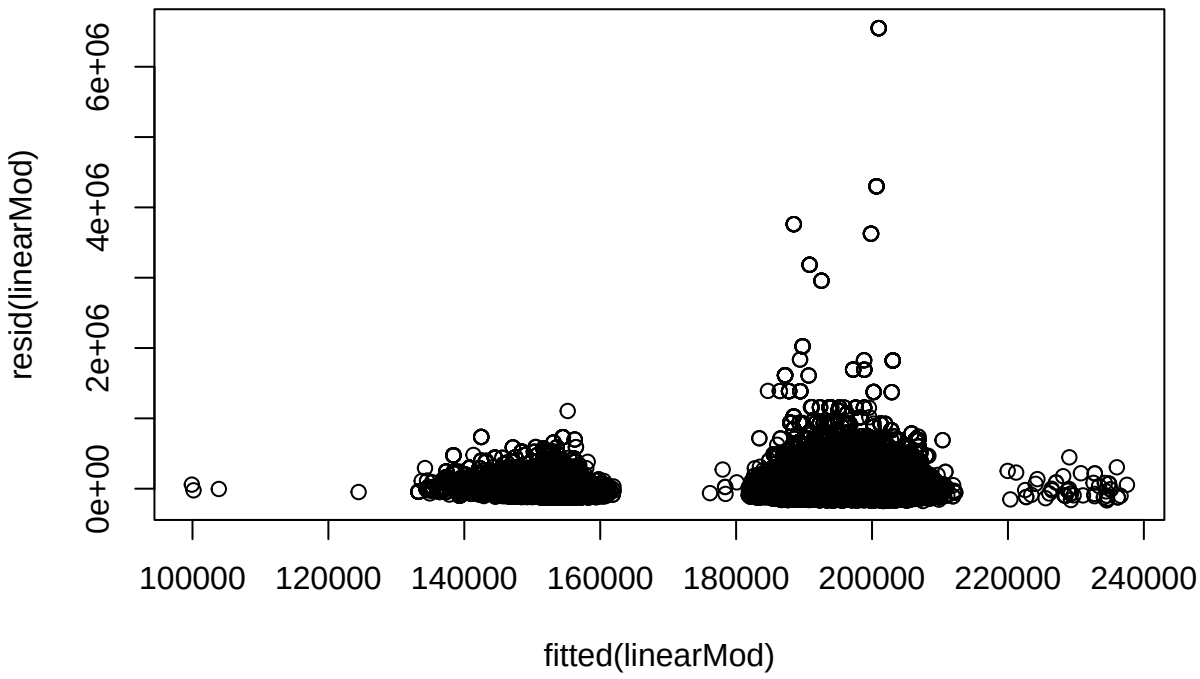
How could you summarize your data to answer key questions?

```
linearMod <- lm(AMT_INCOME_TOTAL ~ STATUS + CNT_CHILDREN +DAYS_BIRTH + DAYS_EMPLOYED, data=card_df1)
summary(linearMod)
```

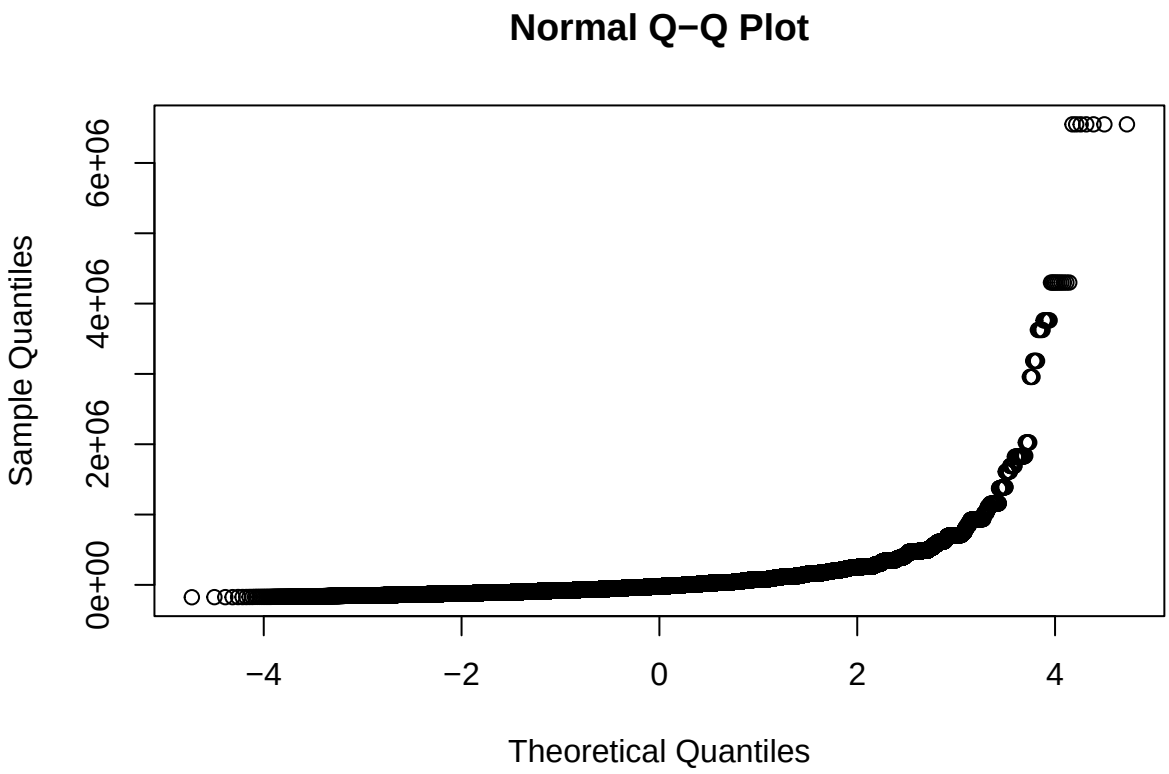
```
##
## Call:
## lm(formula = AMT_INCOME_TOTAL ~ STATUS + CNT_CHILDREN + DAYS_BIRTH +
##     DAYS_EMPLOYED, data = card_df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175987  -63822  -19786   34324 6549013
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.741e+05  8.307e+02 209.573  < 2e-16 ***
## STATUS1        4.863e+03  7.089e+03   0.686  0.49275
## STATUS2       -1.015e+04  2.498e+04  -0.407  0.68436
## STATUS3       -5.366e+04  4.445e+04  -1.207  0.22740
## STATUS4       -3.201e+04  4.869e+04  -0.657  0.51099
## STATUS5        3.410e+04  1.418e+04   2.405  0.01616 *
## STATUSC       -3.108e+03  9.705e+02  -3.202  0.00136 **
## STATUSX        3.573e+03  1.634e+03   2.187  0.02877 *
## CNT_CHILDREN  -7.124e+02  2.422e+02  -2.942  0.00327 **
## DAYS_BIRTH    -1.387e+00  5.177e-02 -26.798  < 2e-16 ***
## DAYS_EMPLOYED -1.388e-01  1.508e-03 -92.075  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108900 on 438546 degrees of freedom
## Multiple R-squared:  0.02185,    Adjusted R-squared:  0.02183
## F-statistic: 979.8 on 10 and 438546 DF,  p-value: < 2.2e-16
```

```
plot(fitted(linearMod), resid(linearMod))
```




```
qqnorm(resid(linearMod))
```



What information is not self-evident

```
logMod <- glm(STATUS ~ AMT_INCOME_TOTAL + CNT_CHILDREN +DAYS_BIRTH + DAYS_EMPLOYED, data=card_df, family
summary(logMod)
```

```
##
## Call:
## glm(formula = STATUS ~ AMT_INCOME_TOTAL + CNT_CHILDREN + DAYS_BIRTH +
##     DAYS_EMPLOYED, family = binomial(link = "logit"), data = card_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2960  -0.2910  -0.2878  -0.2843   2.5949
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.157e+00  4.059e-02 -77.781   <2e-16 ***
## AMT_INCOME_TOTAL -1.076e-07  7.322e-08  -1.470   0.1416
## CNT_CHILDREN     -2.867e-02  1.144e-02  -2.506   0.0122 *
## DAYS_BIRTH       -2.264e-06  2.401e-06  -0.943   0.3456
## DAYS_EMPLOYED    -1.821e-07  7.110e-08  -2.561   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148852  on 438556  degrees of freedom
```

```
## Residual deviance: 148838  on 438552  degrees of freedom
## AIC: 148848
##
## Number of Fisher Scoring iterations: 6
```

```
logMod2 <- glm(CODE_GENDER ~ AMT_INCOME_TOTAL + CNT_CHILDREN +DAYS_BIRTH + DAYS_EMPLOYED, data=card_df,
summary(logMod2)
```

```
##
## Call:
## glm(formula = CODE_GENDER ~ AMT_INCOME_TOTAL + CNT_CHILDREN +
##     DAYS_BIRTH + DAYS_EMPLOYED, family = binomial(link = "logit"),
##     data = card_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4857  -0.9221  -0.7129   1.2723   2.2252
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.330e-01  1.716e-02  -7.751  9.1e-15 ***
## AMT_INCOME_TOTAL  3.559e-06  3.488e-08 102.025  < 2e-16 ***
## CNT_CHILDREN      5.757e-02  4.638e-03  12.411  < 2e-16 ***
## DAYS_BIRTH        7.870e-05  1.036e-06  75.945  < 2e-16 ***
## DAYS_EMPLOYED    -1.198e-06  3.502e-08 -34.195  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 555384  on 438556  degrees of freedom
## Residual deviance: 524391  on 438552  degrees of freedom
## AIC: 524401
##
## Number of Fisher Scoring iterations: 4
```

```
logMod3 <- glm(FLAG_OWN_REALTY ~ AMT_INCOME_TOTAL + CNT_CHILDREN +DAYS_BIRTH + DAYS_EMPLOYED, data=card
summary(logMod3)
```

```
##
## Call:
## glm(formula = FLAG_OWN_REALTY ~ AMT_INCOME_TOTAL + CNT_CHILDREN +
##     DAYS_BIRTH + DAYS_EMPLOYED, family = binomial(link = "logit"),
##     data = card_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3812  -1.4075   0.7837   0.8768   1.1055
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.635e-01  1.733e-02 -20.972  < 2e-16 ***
## AMT_INCOME_TOTAL  7.223e-07  3.417e-08  21.140  < 2e-16 ***
## CNT_CHILDREN      1.299e-01  4.872e-03  26.661  < 2e-16 ***
## DAYS_BIRTH       -6.209e-05  1.034e-06 -60.053  < 2e-16 ***
```

```
## DAYS_EMPLOYED     1.350e-07  3.189e-08   4.235 2.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 540647  on 438556  degrees of freedom
## Residual deviance: 534394  on 438552  degrees of freedom
## AIC: 534404
##
## Number of Fisher Scoring iterations: 4
```

```
logMod4 <- glm(FLAG_OWN_REALTY ~ STATUS + CNT_CHILDREN +DAYS_BIRTH + DAYS_EMPLOYED, data=card_df, family
summary(logMod4)
```

```
##
## Call:
## glm(formula = FLAG_OWN_REALTY ~ STATUS + CNT_CHILDREN + DAYS_BIRTH +
##     DAYS_EMPLOYED, family = binomial(link = "logit"), data = card_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0007  -1.4101   0.7864   0.8753   1.4623
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.303e-01  1.627e-02 -14.154  < 2e-16 ***
## STATUS1       -3.506e-01  1.346e-01  -2.605  0.00918 **
## STATUS2       -6.977e-01  4.643e-01  -1.503  0.13289
## STATUS3       -2.404e-01  8.734e-01  -0.275  0.78312
## STATUS4       -1.080e+00  9.253e-01  -1.167  0.24325
## STATUS5        3.203e-01  3.076e-01   1.041  0.29778
## STATUSC       -1.650e-01  1.896e-02  -8.699  < 2e-16 ***
## STATUSX       -1.991e-01  3.169e-02  -6.284  3.3e-10 ***
## CNT_CHILDREN   1.288e-01  4.868e-03  26.459  < 2e-16 ***
## DAYS_BIRTH    -6.301e-05  1.033e-06 -61.009  < 2e-16 ***
## DAYS_EMPLOYED  3.775e-08  3.156e-08   1.196  0.23171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 540647  on 438556  degrees of freedom
## Residual deviance: 534748  on 438546  degrees of freedom
## AIC: 534770
##
## Number of Fisher Scoring iterations: 4
```

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain

will do in final step

Questions for future steps

which ml techniques is suitable for my model

want to add some more questions based on plotted results