Correlation Analysis in Credit Prediction: An Empirical Study on Income, Employment, and Family Demographics

Avery Holloman

2024-08-11

```
# libraries
library(ggplot2)
library(dplyr)
```

```
##
## 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(reshape2)
library(readxl)
```

```
# Loading my Data
creditpredictions <- read_excel("C:/Users/jacob/OneDrive/Desktop/R Studio Projects 2024/creditpr
edictions.xlsx")
```

```
# Checking the column names to make sure no duplicates
print(names(creditpredictions))
```

```
##  [1] "ID"                 "CODE_GENDER"         "FLAG_OWN_CAR"
##  [4] "FLAG_OWN_REALTY"    "CNT_CHILDREN"        "AMT_INCOME_TOTAL"
##  [7] "NAME_INCOME_TYPE"   "NAME_EDUCATION_TYPE" "NAME_FAMILY_STATUS"
## [10] "NAME_HOUSING_TYPE"  "DAYS_BIRTH"          "DAYS_EMPLOYED"
## [13] "FLAG_MOBIL"         "FLAG_WORK_PHONE"     "FLAG_PHONE"
## [16] "FLAG_EMAIL"         "OCCUPATION_TYPE"     "CNT_FAM_MEMBERS"
```

```
# Next I need to start by Selecting only the numerical columns
numeric_features <- creditpredictions[, c("ID", "AMT_INCOME_TOTAL", "DAYS_BIRTH", "DAYS_EMPLOYE
D", "CNT_CHILDREN", "CNT_FAM_MEMBERS")]
```

```
# The "head" functions shows the first few rows or features or columns
print(head(numeric_features))
```

```
## # A tibble: 6 × 6
##        ID AMT_INCOME_TOTAL DAYS_BIRTH DAYS_EMPLOYED CNT_CHILDREN CNT_FAM_MEMBERS
##     <dbl>            <dbl>      <dbl>         <dbl>        <dbl>            <dbl>
## 1 5008804           427500     -12005         -4542            0                2
## 2 5008805           427500     -12005         -4542            0                2
## 3 5008806           112500     -21474         -1134            0                2
## 4 5008809           270000     -19110         -3051            0                1
## 5 5008810           270000     -19110         -3051            0                1
## 6 5008811           270000     -19110         -3051            0                1
```

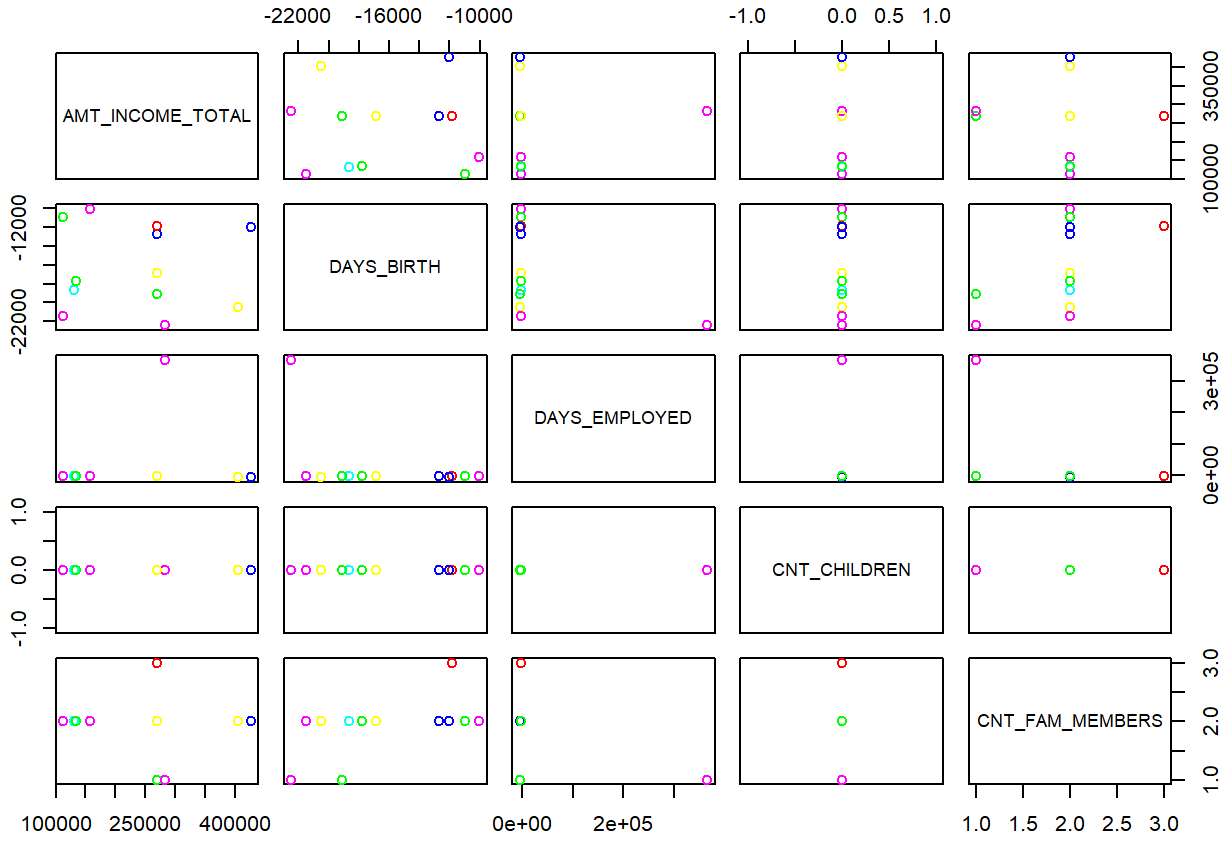
```
# Correlation matrix
correlation_matrix <- cor(numeric_features[, -1], use = "complete.obs")
```

```
## Warning in cor(numeric_features[, -1], use = "complete.obs"): the standard
## deviation is zero
```

```
print(correlation_matrix)
```

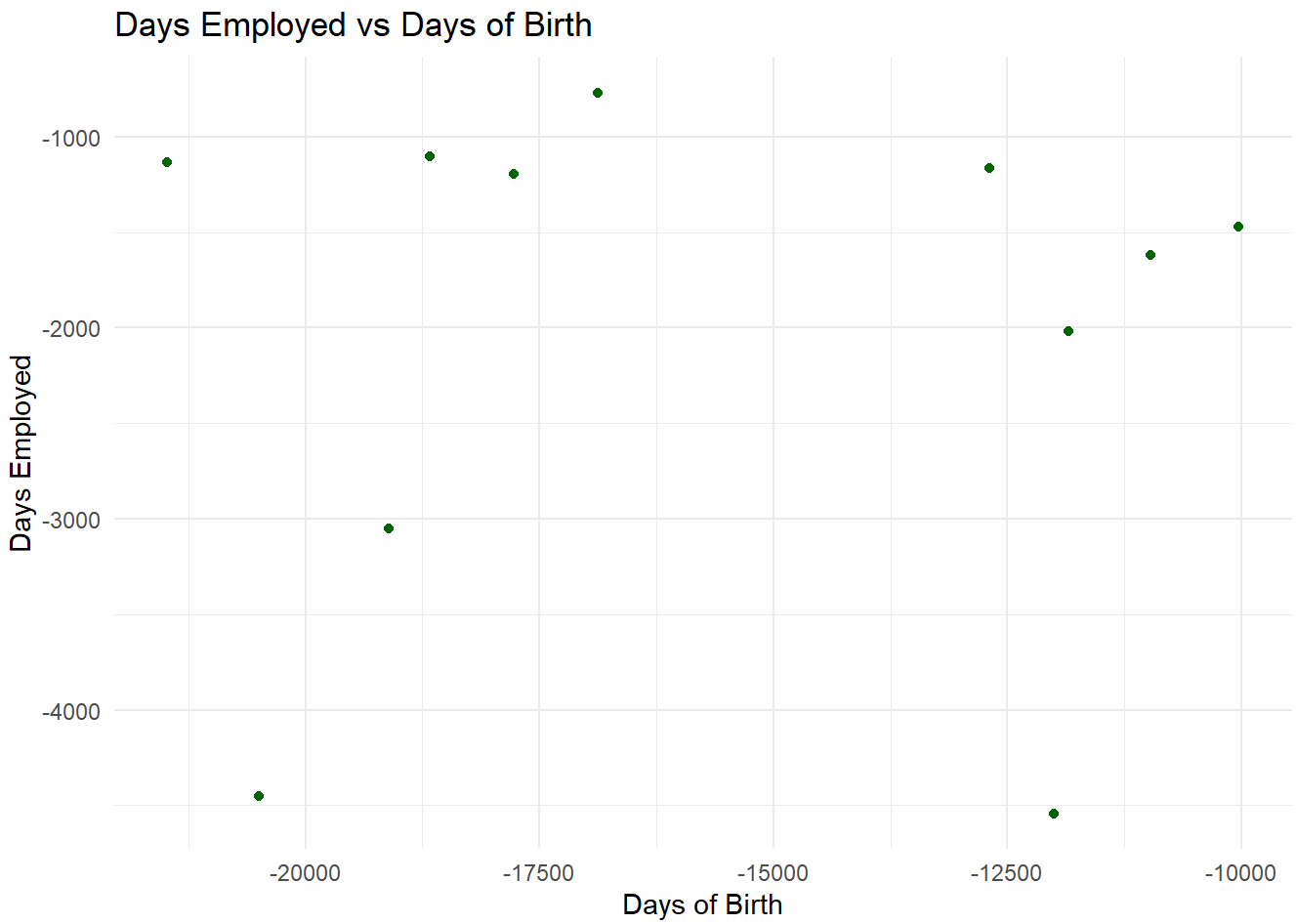
```
##                  AMT_INCOME_TOTAL  DAYS_BIRTH DAYS_EMPLOYED CNT_CHILDREN
## AMT_INCOME_TOTAL       1.00000000 -0.01530798     0.2181920           NA
## DAYS_BIRTH            -0.01530798  1.00000000    -0.5017094           NA
## DAYS_EMPLOYED          0.21819204 -0.50170944     1.0000000           NA
## CNT_CHILDREN                   NA          NA            NA            1
## CNT_FAM_MEMBERS       -0.17151254  0.58518394    -0.5758951           NA
##                  CNT_FAM_MEMBERS
## AMT_INCOME_TOTAL      -0.1715125
## DAYS_BIRTH             0.5851839
## DAYS_EMPLOYED         -0.5758951
## CNT_CHILDREN                  NA
## CNT_FAM_MEMBERS        1.0000000
```

```
# Scatter plot matrix with rainbow colors
pairs(numeric_features[, -1], col = rainbow(6))
```

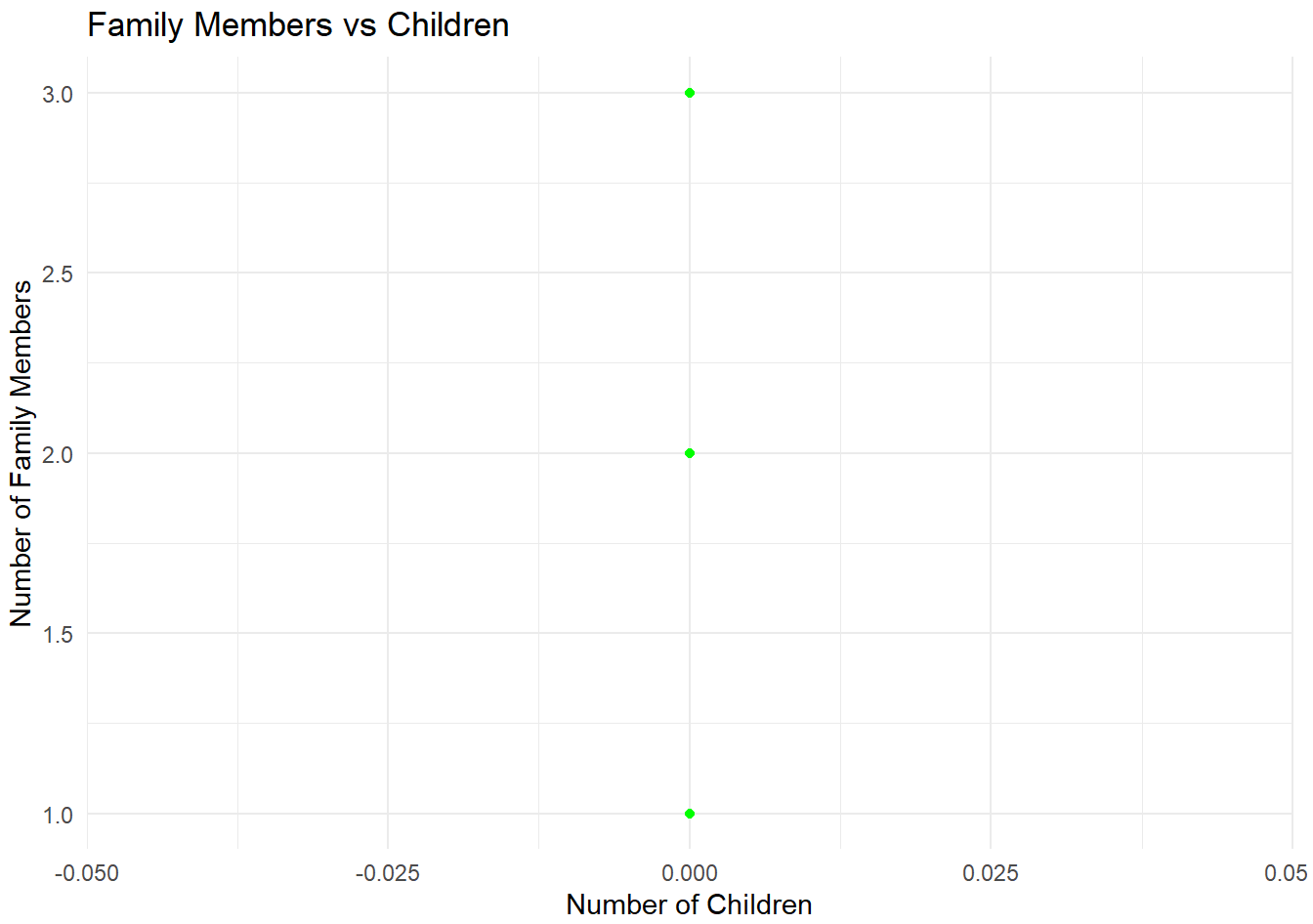



```
# Now I want to see the data with a scatter plot, but I will filter the data for negative employ
ment length and remove unnecessary columns that can cause overfitting
filtered_data <- creditpredictions %>% filter(DAYS_EMPLOYED < 0)
```

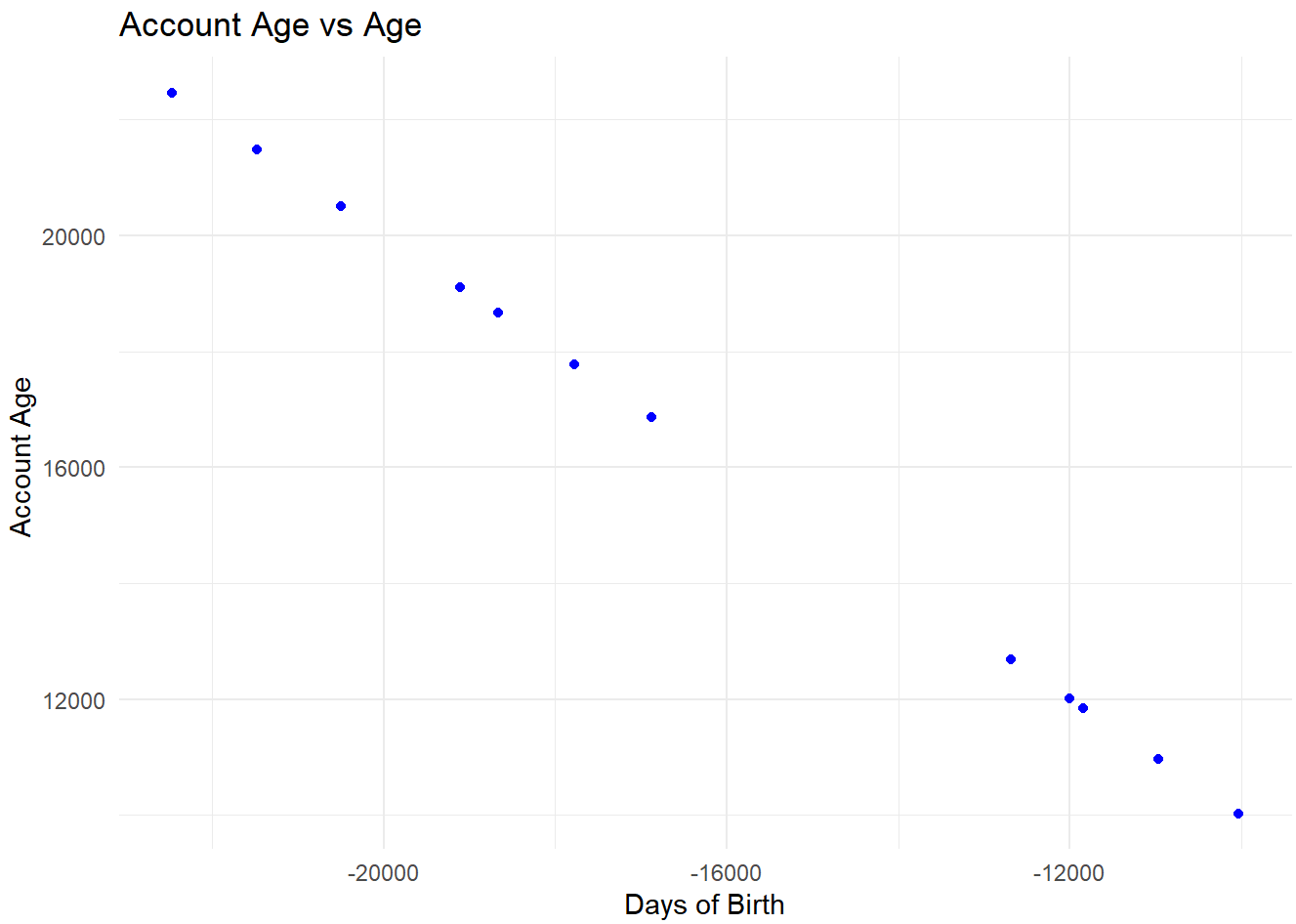
```
ggplot(filtered_data, aes(x = DAYS_BIRTH, y = DAYS_EMPLOYED)) +
  geom_point(alpha = 0.7, color = "darkgreen") +
  theme_minimal() +
  labs(x = "Days of Birth", y = "Days Employed", title = "Days Employed vs Days of Birth")
```




```
# Family member vs children
ggplot(creditpredictions, aes(x = CNT_CHILDREN, y = CNT_FAM_MEMBERS)) +
  geom_point(alpha = 0.7, color = "green") +
  theme_minimal() +
  labs(x = "Number of Children", y = "Number of Family Members", title = "Family Members vs Chil
dren")
```




```
# Account age vs age
ggplot(creditpredictions, aes(x = DAYS_BIRTH, y = -DAYS_BIRTH)) +
  geom_point(alpha = 0.7, color = "blue") +
  theme_minimal() +
  labs(x = "Days of Birth", y = "Account Age ", title = "Account Age vs Age")
```




```
# Employment Length vs age
ggplot(creditpredictions, aes(x = -DAYS_BIRTH, y = -DAYS_EMPLOYED)) +
  geom_point(alpha = 0.7, color = "red") +
  theme_minimal() +
  labs(x = "Age (in days)", y = "Employment Length (in days)", title = "Employment Length vs Ag
e")
```

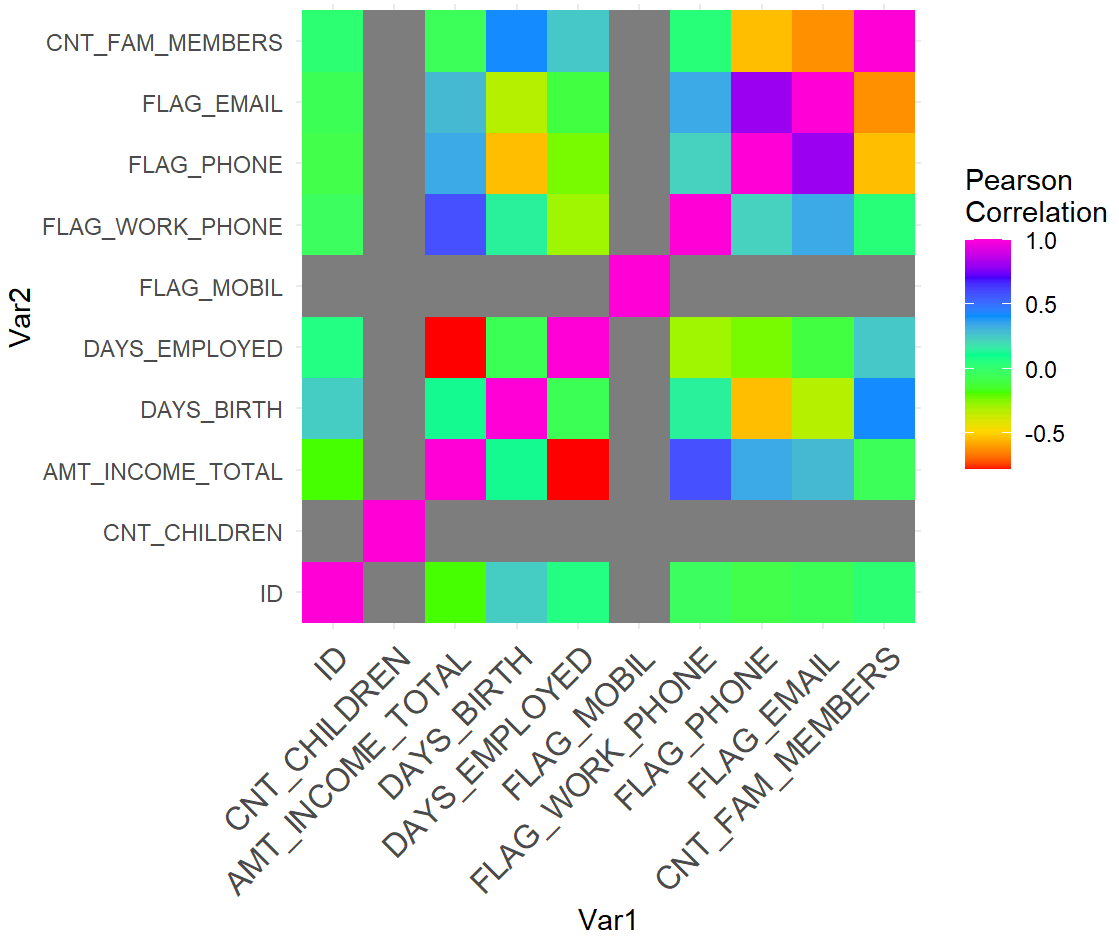

```
# Correlation analysis
correlation_matrix_filtered <- cor(filtered_data %>% select_if(is.numeric), use = "complete.ob
s")
```

```
## Warning in cor(filtered_data %>% select_if(is.numeric), use = "complete.obs"):
## the standard deviation is zero
```

```
print(correlation_matrix_filtered)
```

```
##                          ID CNT_CHILDREN AMT_INCOME_TOTAL  DAYS_BIRTH
## ID                1.00000000           NA      -0.18806417  0.24092040
## CNT_CHILDREN              NA            1               NA          NA
## AMT_INCOME_TOTAL -0.18806417           NA       1.00000000  0.11595001
## DAYS_BIRTH        0.24092040           NA       0.11595001  1.00000000
## DAYS_EMPLOYED     0.06316046           NA      -0.78380163 -0.06007222
## FLAG_MOBIL                NA           NA               NA          NA
## FLAG_WORK_PHONE  -0.04261869           NA       0.58614281  0.13619858
## FLAG_PHONE       -0.09136256           NA       0.33516595 -0.53848977
## FLAG_EMAIL       -0.06032043           NA       0.28525864 -0.33110040
## CNT_FAM_MEMBERS   0.01936053           NA      -0.05177469  0.41847567
##                  DAYS_EMPLOYED FLAG_MOBIL FLAG_WORK_PHONE  FLAG_PHONE
## ID                  0.06316046         NA     -0.04261869 -0.09136256
## CNT_CHILDREN                NA         NA               NA          NA
## AMT_INCOME_TOTAL   -0.78380163         NA      0.58614281  0.33516595
## DAYS_BIRTH         -0.06007222         NA      0.13619858 -0.53848977
## DAYS_EMPLOYED       1.00000000         NA     -0.29148929 -0.23816941
## FLAG_MOBIL                  NA          1               NA          NA
## FLAG_WORK_PHONE    -0.29148929         NA      1.00000000  0.22174461
## FLAG_PHONE         -0.23816941         NA      0.22174461  1.00000000
## FLAG_EMAIL         -0.11151269         NA      0.33265142  0.80423994
## CNT_FAM_MEMBERS     0.25132342         NA      0.03784579 -0.54234111
##                   FLAG_EMAIL CNT_FAM_MEMBERS
## ID                -0.06032043      0.01936053
## CNT_CHILDREN               NA              NA
## AMT_INCOME_TOTAL   0.28525864     -0.05177469
## DAYS_BIRTH        -0.33110040      0.41847567
## DAYS_EMPLOYED     -0.11151269      0.25132342
## FLAG_MOBIL                 NA              NA
## FLAG_WORK_PHONE    0.33265142      0.03784579
## FLAG_PHONE         0.80423994     -0.54234111
## FLAG_EMAIL         1.00000000     -0.62743875
## CNT_FAM_MEMBERS   -0.62743875      1.00000000
```

```
# Correlation heatmap
melted_corr <- melt(correlation_matrix_filtered)
ggplot(melted_corr, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradientn(colors = rainbow(7), name="Pearson\nCorrelation") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1)) +
coord_fixed()
```




```
# Numerical vs categorical (ANOVA)
anova_results <- aov(AMT_INCOME_TOTAL ~ NAME_INCOME_TYPE + NAME_EDUCATION_TYPE + NAME_FAMILY_STA
TUS + NAME_HOUSING_TYPE, data = creditpredictions)

summary(anova_results)
```

```
##                        Df    Sum Sq   Mean Sq F value Pr(>F)
## NAME_INCOME_TYPE        2 5.042e+12 2.521e+12   902.9 <2e-16 ***
## NAME_EDUCATION_TYPE     2 5.339e+13 2.670e+13  9561.6 <2e-16 ***
## NAME_FAMILY_STATUS      2 1.126e+13 5.631e+12  2016.7 <2e-16 ***
## Residuals           10532 2.941e+13 2.792e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Age vs all categorical columns not included
anova_age_vs_cat <- aov(-DAYS_BIRTH ~ NAME_INCOME_TYPE + NAME_EDUCATION_TYPE + NAME_FAMILY_STATU
S + NAME_HOUSING_TYPE, data = creditpredictions)

summary(anova_age_vs_cat)
```

```
##                         Df    Sum Sq   Mean Sq F value Pr(>F)
## NAME_INCOME_TYPE         2 7.821e+10 3.911e+10  5067.7 <2e-16 ***
## NAME_EDUCATION_TYPE      2 2.000e+10 9.999e+09  1295.8 <2e-16 ***
## NAME_FAMILY_STATUS       2 1.563e+09 7.815e+08   101.3 <2e-16 ***
## Residuals            10532 8.127e+10 7.717e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Categorical vs categorical
# Chi-square test verus NAME_INCOME_TYPE

chi_square_results <- function(feature){
observed <- table(creditpredictions[[feature]],      creditpredictions$NAME_INCOME_TYPE)
expected <- chisq.test(observed)$expected
test_result <- chisq.test(observed)
  list(observed = observed,expected = expected,
chi_square = test_result$statistic,
critical_value = qchisq(0.95, df = (nrow(observed) - 1) *  (ncol(observed) - 1)),
    p_value = test_result$p.value)
}
```

```
# Next I will select specific categorical columns
cat_features <- c("CODE_GENDER", "FLAG_OWN_CAR", "FLAG_OWN_REALTY", "NAME_EDUCATION_TYPE", "NAME
_FAMILY_STATUS", "NAME_HOUSING_TYPE")

chi_square_results_list <- lapply(cat_features, chi_square_results)

# Lets review the results
for (i in seq_along(cat_features)) {
  cat("\nChi-square test for:", cat_features[i], "\n")
  print(chi_square_results_list[[i]])
}
```

```
##
## Chi-square test for: CODE_GENDER
## $observed
##
##     Commercial associate Pensioner Working
##   F                 1092       364    2541
##   M                 3268       728    2546
##
## $expected
##
##     Commercial associate Pensioner  Working
##   F             1653.565  414.1497 1929.285
##   M             2706.435  677.8503 3157.715
##
## $chi_square
## X-squared
##  629.4725
##
## $critical_value
## [1] 5.991465
##
## $p_value
## [1] 2.05014e-137
##
##
## Chi-square test for: FLAG_OWN_CAR
## $observed
##
##     Commercial associate Pensioner Working
##   N                 1092      1092    2541
##   Y                 3268         0    2546
##
## $expected
##
##     Commercial associate Pensioner  Working
##   N              1954.74  489.5816 2280.679
##   Y              2405.26  602.4184 2806.321
##
## $chi_square
## X-squared
##  2087.773
##
## $critical_value
## [1] 5.991465
##
## $p_value
## [1] 0
##
##
## Chi-square test for: FLAG_OWN_REALTY
## $observed
##
```

```
##     Commercial associate Pensioner Working
##   Y                 4360      1092    5087
##
## $expected
## [1] 3513 3513 3513
##
## $chi_square
## X-squared
##   2577.89
##
## $critical_value
## [1] 0
##
## $p_value
## [1] 0
##
##
## Chi-square test for: NAME_EDUCATION_TYPE
## $observed
##
##                                Commercial associate Pensioner Working
##   Higher education                             1089      1092    1456
##   Incomplete higher                               0         0     726
##   Secondary / secondary special                3271         0    2905
##
## $expected
##
##                                Commercial associate Pensioner   Working
##   Higher education                         1504.6323 376.84828 1755.5194
##   Incomplete higher                         300.3473  75.22459  350.4281
##   Secondary / secondary special            2555.0204 639.92713 2981.0525
##
## $chi_square
## X-squared
##  3143.665
##
## $critical_value
## [1] 9.487729
##
## $p_value
## [1] 0
##
##
## Chi-square test for: NAME_FAMILY_STATUS
## $observed
##
##                        Commercial associate Pensioner Working
##   Civil marriage                          0         0     728
##   Married                              3268         0    3270
##   Separated                               0      1092       0
##   Single / not married                 1092         0    1089
##
```

```
## $expected
##
##                         Commercial associate Pensioner   Working
##   Civil marriage                   301.1747  75.43182  351.3935
##   Married                         2704.7803 677.43581 3155.7839
##   Separated                        451.7620 113.14774  527.0902
##   Single / not married             902.2829 225.98463 1052.7324
##
## $chi_square
## X-squared
##  11293.21
##
## $critical_value
## [1] 12.59159
##
## $p_value
## [1] 0
##
##
## Chi-square test for: NAME_HOUSING_TYPE
## $observed
##
##                      Commercial associate Pensioner Working
##   House / apartment                  4360      1092    4359
##   Rented apartment                      0         0     728
##
## $expected
##
##                      Commercial associate  Pensioner   Working
##   House / apartment            4058.8253 1016.56818 4735.6065
##   Rented apartment              301.1747   75.43182  351.3935
##
## $chi_square
## X-squared
##  838.1304
##
## $critical_value
## [1] 5.991465
##
## $p_value
## [1] 1.005277e-182
```