FinalProject

Blanchard

3/23/2020

We will be looking at the Federal Home Loan Housing Banks 2018 dataset. This is a public database of all transactions made by the 11 FHLH banks in the United States. A quick bit of information, most loan criteria is classified based on the Property type, borrower income and credit score, LTV or Loan to Value ratio, DTI or debt to income, and the purpose of the loan (rate and term refinance, purchase, cashout refinance, etc. Some of the previosly mentioned criteria is included in the dataset, however my initial concern is that because all of the loans were made, and most criteria falls into a similar grouping, like all credit scores should be above 500, and all LTV ratios are all below 100%, it may be difficult to find a linear regression correlation between the data. But we shall try. Below is our summary:

```
Bank <- read.csv(file.choose()) #MovieRating.csv head(Bank)
summary(Bank)
```

```
##       Year        AssignedID                Bank       FIPSStateCode
##  Min.   :2018   Min.   :2128273   Chicago     :12994   Min.   : 1.00
##  1st Qu.:2018   1st Qu.:2144699   Topeka      : 9577   1st Qu.:18.00
##  Median :2018   Median :2161124   Indianapolis: 9514   Median :26.00
##  Mean   :2018   Mean   :2161124   Cincinnati  : 8955   Mean   :28.49
##  3rd Qu.:2018   3rd Qu.:2177550   Des Moines  : 8738   3rd Qu.:39.00
##  Max.   :2018   Max.   :2193975   Dallas      : 5269   Max.   :72.00
##                                   (Other)     :10656
##  FIPSCountyCode        MSA          FeatureID             Tract
##  Min.   :  1.00   Min.   :10100   Min.   :6.920e+03   Min.   :   1.0
##  1st Qu.: 31.00   1st Qu.:19220   1st Qu.:2.411e+06   1st Qu.: 103.4
##  Median : 71.00   Median :30780   Median :1.000e+10   Median : 508.0
##  Mean   : 87.95   Mean   :34979   Mean   :7.348e+09   Mean   :2774.7
##  3rd Qu.:119.00   3rd Qu.:41180   3rd Qu.:1.000e+10   3rd Qu.:5105.0
##  Max.   :840.00   Max.   :99999   Max.   :1.000e+10   Max.   :9900.0
##
##      MinPer          TraMedY          LocMedY          Tractrat
##  Min.   :  0.00   Min.   :     0   Min.   : 18666   Min.   :0.0000
##  1st Qu.:  5.95   1st Qu.: 62375   1st Qu.: 62393   1st Qu.:0.9341
##  Median : 12.54   Median : 77318   Median : 69434   Median :1.1100
##  Mean   : 19.54   Mean   : 83313   Mean   : 70535   Mean   :1.1717
##  3rd Qu.: 25.62   3rd Qu.: 99720   3rd Qu.: 75898   3rd Qu.:1.3600
##  Max.   :100.00   Max.   :250001   Max.   :127500   Max.   :3.9037
##
##      Income           CurAreY           IncRat              UPB
##  Min.   :  11004   Min.   : 18600   Min.   :  0.1586   Min.   :  10400
##  1st Qu.:  63492   1st Qu.: 65800   1st Qu.:  0.8651   1st Qu.: 139650
##  Median :  95000   Median : 73600   Median :  1.2746   Median : 210500
##  Mean   : 112689   Mean   : 75020   Mean   :  1.5221   Mean   : 237381
##  3rd Qu.: 137035   3rd Qu.: 80600   3rd Qu.:  1.8258   3rd Qu.: 306507
##  Max.   :9614088   Max.   :134800   Max.   :148.2309   Max.   :1190000
##
##       LTV           MortDate       AcquDate       Purpose           Coop
##  Min.   :0.008   Min.   :2016   Min.   :2018   Min.   :1.000   Min.   :2
##  1st Qu.:0.720   1st Qu.:2018   1st Qu.:2018   1st Qu.:1.000   1st Qu.:2
##  Median :0.800   Median :2018   Median :2018   Median :1.000   Median :2
##  Mean   :0.771   Mean   :2018   Mean   :2018   Mean   :1.341   Mean   :2
##  3rd Qu.:0.850   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2.000   3rd Qu.:2
##  Max.   :1.680   Max.   :2018   Max.   :2018   Max.   :4.000   Max.   :2
##
##     Product     FedGuar            Term          AmorTerm        SellType
##  Min.   :1   Min.   :0.0000   Min.   : 60.0   Min.   : 60.0   Min.   :1
##  1st Qu.:1   1st Qu.:0.0000   1st Qu.:360.0   1st Qu.:360.0   1st Qu.:1
##  Median :1   Median :0.0000   Median :360.0   Median :360.0   Median :1
##  Mean   :1   Mean   :0.1355   Mean   :344.3   Mean   :344.3   Mean   :1
##  3rd Qu.:1   3rd Qu.:0.0000   3rd Qu.:360.0   3rd Qu.:360.0   3rd Qu.:1
##  Max.   :1   Max.   :3.0000   Max.   :360.0   Max.   :360.0   Max.   :1
##
##      NumBor          First            CICA           BoRace
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:9.000   1st Qu.:5.000
##  Median :2.000   Median :2.000   Median :9.000   Median :5.000
```

```
##  Mean   :1.561   Mean   :1.771   Mean   :8.976   Mean   :5.014
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:9.000   3rd Qu.:5.000
##  Max.   :4.000   Max.   :2.000   Max.   :9.000   Max.   :7.000
##
##      CoRace          BoGender        CoGender         BoAge           CoAge
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :18.0   Min.   :18.00
##  1st Qu.:5.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:34.0   1st Qu.:40.00
##  Median :7.000   Median :1.000   Median :2.000   Median :44.0   Median :81.00
##  Mean   :6.374   Mean   :1.402   Mean   :2.832   Mean   :48.7   Mean   :70.62
##  3rd Qu.:8.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:58.0   3rd Qu.:98.00
##  Max.   :8.000   Max.   :3.000   Max.   :4.000   Max.   :99.0   Max.   :99.00
##
##      Occup           NumUnits          Bed1         Bed2         Bed3
##  Min.   :1.000   Min.   :1.000   Min.   :98   Min.   :98   Min.   :98
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:98   1st Qu.:98   1st Qu.:98
##  Median :1.000   Median :1.000   Median :98   Median :98   Median :98
##  Mean   :1.032   Mean   :1.011   Mean   :98   Mean   :98   Mean   :98
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:98   3rd Qu.:98   3rd Qu.:98
##  Max.   :2.000   Max.   :4.000   Max.   :98   Max.   :98   Max.   :98
##
##       Bed4         Aff1        Aff2        Aff3        Aff4       Rent1
##  Min.   :98   Min.   :2   Min.   :2   Min.   :2   Min.   :2   Min.   :1e+10
##  1st Qu.:98   1st Qu.:2   1st Qu.:2   1st Qu.:2   1st Qu.:2   1st Qu.:1e+10
##  Median :98   Median :2   Median :2   Median :2   Median :2   Median :1e+10
##  Mean   :98   Mean   :2   Mean   :2   Mean   :2   Mean   :2   Mean   :1e+10
##  3rd Qu.:98   3rd Qu.:2   3rd Qu.:2   3rd Qu.:2   3rd Qu.:2   3rd Qu.:1e+10
##  Max.   :98   Max.   :2   Max.   :2   Max.   :2   Max.   :2   Max.   :1e+10
##
##      Rent2           Rent3           Rent4          RentUt1
##  Min.   :1e+10   Min.   :1e+10   Min.   :1e+10   Min.   :1e+10
##  1st Qu.:1e+10   1st Qu.:1e+10   1st Qu.:1e+10   1st Qu.:1e+10
##  Median :1e+10   Median :1e+10   Median :1e+10   Median :1e+10
##  Mean   :1e+10   Mean   :1e+10   Mean   :1e+10   Mean   :1e+10
##  3rd Qu.:1e+10   3rd Qu.:1e+10   3rd Qu.:1e+10   3rd Qu.:1e+10
##  Max.   :1e+10   Max.   :1e+10   Max.   :1e+10   Max.   :1e+10
##
##     RentUt2         RentUt3         RentUt4           Geog
##  Min.   :1e+10   Min.   :1e+10   Min.   :1e+10   Min.   :1.000
##  1st Qu.:1e+10   1st Qu.:1e+10   1st Qu.:1e+10   1st Qu.:2.000
##  Median :1e+10   Median :1e+10   Median :1e+10   Median :2.000
##  Mean   :1e+10   Mean   :1e+10   Mean   :1e+10   Mean   :1.967
##  3rd Qu.:1e+10   3rd Qu.:1e+10   3rd Qu.:1e+10   3rd Qu.:2.000
##  Max.   :1e+10   Max.   :1e+10   Max.   :1e+10   Max.   :2.000
##
##       Rate             Amount            Front             Back
##  Min.   :0.02750   Min.   :  10400   Min.   :0.0001   Min.   :0.0100
##  1st Qu.:0.04380   1st Qu.: 139920   1st Qu.:0.1506   1st Qu.:0.2584
##  Median :0.04630   Median : 211000   Median :0.2011   Median :0.3300
##  Mean   :0.04555   Mean   : 237563   Mean   :0.2143   Mean   :0.3268
##  3rd Qu.:0.04750   3rd Qu.: 307000   3rd Qu.:0.2659   3rd Qu.:0.4000
##  Max.   :0.05880   Max.   :1190000   Max.   :9.9900   Max.   :0.9640
##
```

```
##  BoCreditScore   CoBoCreditScore      PMI               Self
##  Min.   :1.000   Min.   :1.000   Min.   :0.00000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:5.000   1st Qu.:0.00000   1st Qu.:2.000
##  Median :5.000   Median :5.000   Median :0.00000   Median :2.000
##  Mean   :4.342   Mean   :6.486   Mean   :0.05192   Mean   :1.892
##  3rd Qu.:5.000   3rd Qu.:9.000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :0.95000   Max.   :2.000
##
##     PropType        ArmIndex     ArmMarg            PrepayP
##  PT01   :48473   Min.   :99   Min.   :99999   12/31/9999:65703
##  PT07   :10096   1st Qu.:99   1st Qu.:99999
##  PT11   : 2937   Median :99   Median :99999
##  PT02   : 2456   Mean   :99   Mean   :99999
##  PT06   : 1034   3rd Qu.:99   3rd Qu.:99999
##  PT04   :  509   Max.   :99   Max.   :99999
##  (Other):  198
##      BoEth           Race2           Race3       Race4       Race5
##  Min.   :1.000   Min.   :1.000   Min.   :1   Min.   :4   Min.   :5
##  1st Qu.:2.000   1st Qu.:6.000   1st Qu.:6   1st Qu.:6   1st Qu.:6
##  Median :2.000   Median :6.000   Median :6   Median :6   Median :6
##  Mean   :2.067   Mean   :5.993   Mean   :6   Mean   :6   Mean   :6
##  3rd Qu.:2.000   3rd Qu.:6.000   3rd Qu.:6   3rd Qu.:6   3rd Qu.:6
##  Max.   :3.000   Max.   :6.000   Max.   :6   Max.   :6   Max.   :6
##
##      CoEth          Corace2         Corace3         Corace4
##  Min.   :1.000   Min.   :1.000   Min.   :3.000   Min.   :4.000
##  1st Qu.:2.000   1st Qu.:6.000   1st Qu.:6.000   1st Qu.:6.000
##  Median :3.000   Median :6.000   Median :6.000   Median :6.000
##  Mean   :3.386   Mean   :6.892   Mean   :6.894   Mean   :6.894
##  3rd Qu.:5.000   3rd Qu.:8.000   3rd Qu.:8.000   3rd Qu.:8.000
##  Max.   :5.000   Max.   :8.000   Max.   :8.000   Max.   :8.000
##
##     Corace5          HOEPA     LienStatus  SpcHsgGoals   FedFinStbltyPlan
##  Min.   :5.000   Min.   :2   Min.   :1    Min.   :1.00   Min.   :2
##  1st Qu.:6.000   1st Qu.:2   1st Qu.:1    1st Qu.:2.00   1st Qu.:2
##  Median :6.000   Median :2   Median :1    Median :2.00   Median :2
##  Mean   :6.894   Mean   :2   Mean   :1    Mean   :1.97   Mean   :2
##  3rd Qu.:8.000   3rd Qu.:2   3rd Qu.:1    3rd Qu.:2.00   3rd Qu.:2
##  Max.   :8.000   Max.   :2   Max.   :1    Max.   :2.00   Max.   :2
##
##      AcqTyp          GSEREO
##  Min.   :1.000   Min.   :2
##  1st Qu.:1.000   1st Qu.:2
##  Median :1.000   Median :2
##  Mean   :1.409   Mean   :2
##  3rd Qu.:1.000   3rd Qu.:2
##  Max.   :4.000   Max.   :2
##
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
#All information below is the meaning of the variables found in each of the data set columns. I unfortunately was unsuccessful in renaming some of the data. I tried downloading a library calle d dplyr for a function called mutate but kept getting issues regarding it was a field made up of characters. My original goals was update the variables to their alphabetic values rather than th eir numeric groupings within the file.
```

AMA_PUDB_definitions_2018.xlsx

Data Definitions

Federal Housing Finance Agency

Federal Home Loan Bank Purchased Mortgage Files

For 2018 Data Release

Effective 10/18/16

	Column	Definition
1	Year	Year Loan was reported
2	Assigned ID	Unique Record ID (Not Actual Loan Number)
3	Bank	Name of Federal Home Loan Bank District
4	FIPSStateCode	FIPS State Code
5	FIPSCountyCode	FIPS County Code
6	MSA	Core Based Statistical Area 99999=State/County/Tract combination is not located in a CBSA
7	FeatureID	Geographic Names Information System (GNIS) Feature ID
8	Tract	Census Tract
9	MinPer	Census Tract Minority Percentage
10	TraMedY	Census Tract Median Family Income in Whole Dollars
11	LocMedY	Local Area Median Income in Whole Dollars
12	TractRat	Tract Income Ratio
13	Income	Total Borrower(s) Annual Income in Whole Dollars
14	CurAreY	Area Median Family Income in Whole Dollars
15	IncRat	Borrower Income Ratio
16	UPB	Acquisition Unpaid Principal Balance in Whole Dollars
17	LTV	Loan to Value Ratio at Origination
18	MortDate	Year of Mortgage Note
19	AcquDate	Year of Acquisition Note
20	Purpose	Loan Purpose 1=purchase; 2=refinancing; 3=second mortgage; 4=new construction; 5=rehabilitation
21	Coop	Cooperative Unit Mortgage 1=yes; 2=no
22	Product	Product Type 01=Fixed Rate; 02=ARM; 03=Balloon; 04=GPM/GEM; 05=Reverse Annuity Mortgage; 06=other
23	FedGuar	Federal Guarantee 0=no federal guarantee; 1=FHA; 2=VA; 3=FMHA-Guaranteed Rural Housing Loan; 4=HECMs; 5=Title1-FHA
24	Term	Term of Mortgage at Origination in months
25	AmorTerm	Amortization Term in months
26	SellType	Type of Acquiring Lender Institution 1=Insured depository institution; 2=Housing Associate; 3=Insurance Company
27	NumBor	Number of Borrowers
28	First	First Time Home Buyer 1=yes; 2=no

Field Meanings

	Column	Definition
29	CICA	Code indicating whether the mortgage is on a project funded under an AHP, CIP or other CICA program. 1=AHP; 2=CIP; 3=Other CICA 9=Not Applicable; Not Purchased under AHP, CIP, or CICA
30	BoRace	Borrower Race or National Origin 1=American Indian or Alaskan Native; 2=Asian; 3=Black or African American; 4=Native Hawaiian or Other Pacific Islander; 5=White; 7=Information not provided by applicant in mail or telephone application
31	CoRace	Co-borrower Race or National Origin 1=American Indian or Alaskan Native; 2=Asian; 3=Black or African American; 4=Native Hawaiian or Other Pacific Islander; 5=White; 7=Information not provided by applicant in mail or telephone application; 8=No Co-Borrower
32	BoGender	Borrower Gender 1=Male; 2=Female; 3=Information not provided by applicant in mail or telephone application
33	CoGender	Co-Borrower Gender 1=Male; 2=Female; 3=Information not provided by applicant in mail or telephone application; 4=No Co-Borrower
34	BoAge	Age of Borrower in years 99=Not provided by applicant in mail or telephone application
35	CoAge	Age of Co-Borrower in years 98=No Co-Borrower; 99=Not provided by applicant in mail or telephone application
36	Occup	Occupancy Code 1=Principal residence/owner-occupied; 2=second home; 3=investment property (rental)
37	NumUnits	Number of Units
38	Bed1	Unit1--Number of Bedrooms 98=no non-owner-occupied dwelling units
39	Bed2	Unit2--Number of Bedrooms 98=no non-owner-occupied dwelling units
40	Bed3	Unit3--Number of Bedrooms 98=no non-owner-occupied dwelling units
41	Bed4	Unit4--Number of Bedrooms 98=no non-owner-occupied dwelling units
42	Aff1	Unit1--Affordable Category meets the housing goals implemented by HERA Section 1205 1=yes; 2=no
43	Aff2	Unit2--Affordable Category meets the housing goals implemented by HERA Section 1205 1=yes; 2=no
44	Aff3	Unit3--Affordable Category meets the housing goals implemented by HERA Section 1205 1=yes; 2=no

Field Meanings

	Column	Definition
45	Aff4	Unit4--Affordable Category meets the housing goals implemented by HERA Section 1205 1=yes; 2=no
46	Rent1	Unit1-- Reported Rent Level in whole dollars 9999999999=no non-owner-occupied dwelling units
47	Rent2	Unit2-- Reported Rent Level in whole dollars 9999999999=no non-owner-occupied dwelling units
48	Rent3	Unit3-- Reported Rent Level in whole dollars 9999999999=no non-owner-occupied dwelling units
49	Rent4	Unit4-- Reported Rent Level in whole dollars 9999999999=no non-owner-occupied dwelling units
50	RentUt1	Unit1--Reported Rent Plus Utilities in whole dollars 9999999999=no non-owner-occupied dwelling units
51	RentUt2	Unit2--Reported Rent Plus Utilities in whole dollars 9999999999=no non-owner-occupied dwelling units
52	RentUt3	Unit3--Reported Rent Plus Utilities in whole dollars 9999999999=no non-owner-occupied dwelling units
53	RentUt4	Unit4--Reported Rent Plus Utilities in whole dollars 9999999999=no non-owner-occupied dwelling units
54	Geog	Geographically Targeted Indicator indicating whether the property is in census tracts annually classified as underserved by HUD 1=yes; 2=no
55	Rate	Interest Rate
56	Amount	Loan Amount in Whole Dollars
57	Front	Front-end Ratio
58	Back	Back-end Ratio
59	BoCreditScor	Credit Scores are separated into ranges: 1 = <620, 2 = 620 to < 660, 3 = 660 < 700, 4 = 700 < 760, 5 = 760 or greater 9=missing
60	CoBoCreditScor	Credit Scores are separated into ranges: 1 = <620, 2 = 620 to < 660, 3 = 660 < 700, 4 = 700 < 760, 5 = 760 or greater, 9 = missing or no co-borrower
61	PMI	Percent of the original loan balance covered by primary mortgage insurance (PMI) 0=no PMI
62	Self	Self-Employed Indicator 1=yes; 2=no
63	PropType	Property Type PT01=Single family detached; PT02=Deminimus PUD; PT03=Single Family Attached; PT04=Two family; PT05=Townhouse; PT06=Low-Rise Condominium; PT07=PUD; PT08=Duplex; PT09=three family; PT10=Four family; PT11=Hi-Rise condominium; PT12=Manufactured Home
64	ArmIndex	ARM Index 01=11th District Cost of Funds; 02=Other Cost of Funds Index; 03=LIBOR; 04=1 year Constant Treasury Rate 99=Not an ARM
65	ArmMarg	ARM Margin 99999=Not an ARM
66	PrepayP	Prepayment Penalty Date 12/31/9999=No Prepayment Penalty Date

Field Meanings

	Column	Definition
67	BoEth	Borrower Ethnicity 1 -- Hispanic or Latino 2 -- Not Hispanic or Latino 3 -- Information not provided by borrower in mail, Internet, or telephone application
68	Race2	Borrower Race 2 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application
69	Race3	Borrower Race 3 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application
70	Race4	Borrower Race 4 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application
71	Race5	Borrower Race 5 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application
72	CoEth	Co-Borrower Ethnicity 1 -- Hispanic or Latino 2 -- Not Hispanic or Latino 3 -- Information not provided by borrower in mail, Internet, or telephone application 5 -- No Co-Borrower
73	CoRace2	Co-Borrower Race 2 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application 8 -- No Co-borrower
74	CoRace3	Co-Borrower Race 3 1 -- American Indian or Alaska Native 2 -- Asian 3 -- Black or African American 4 -- Native Hawaiian or Other Pacific Islander 5 -- White 6 -- Information not provided by borrower in mail, Internet, or telephone application 8 -- No Co-borrower

Field Meanings

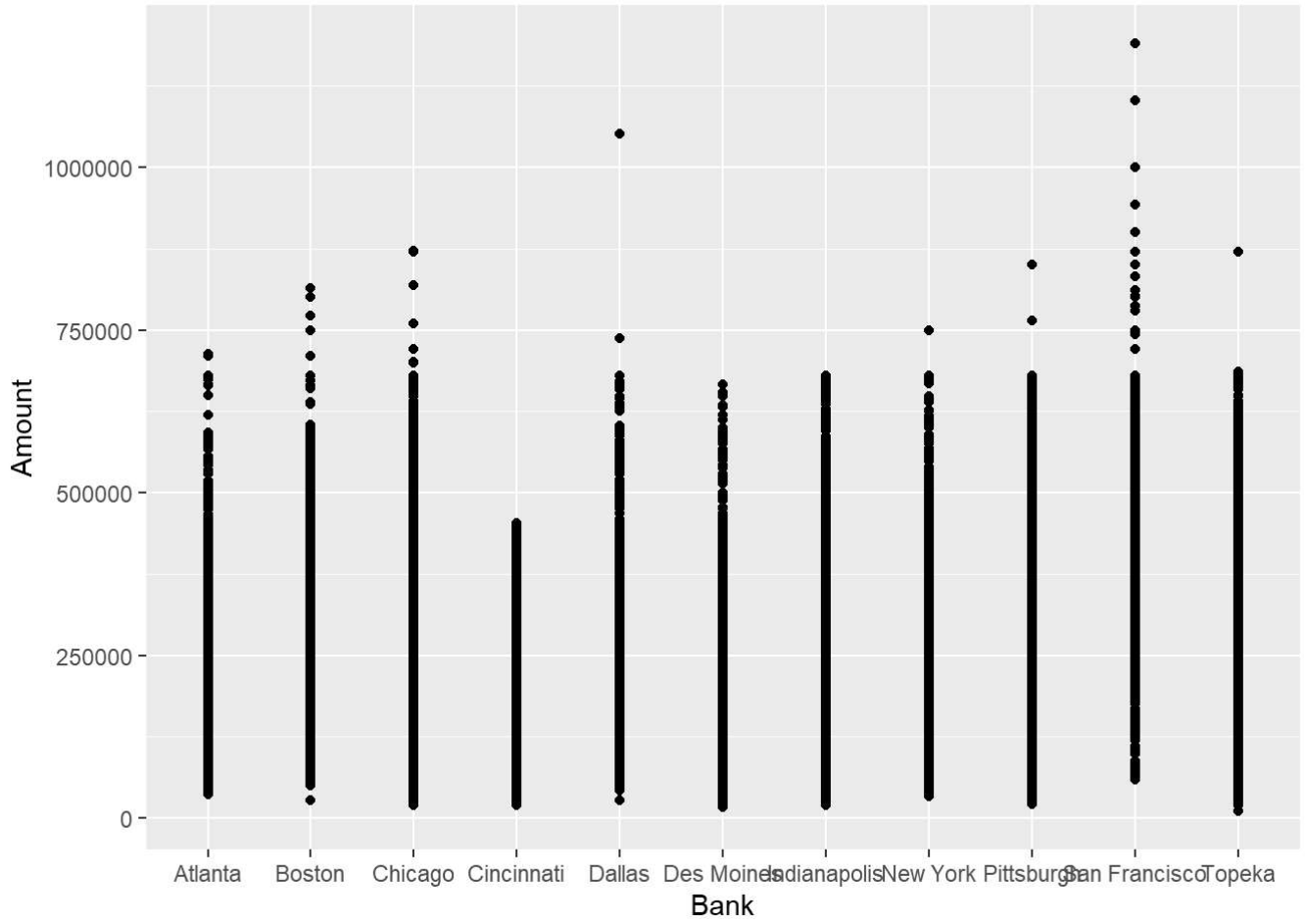
AMA_PUDB_definitions_2018.xlsx

	Column	Definition
75	CoRace4	Co-Borrower Race 4 1 – American Indian or Alaska Native 2 – Asian 3 – Black or African American 4 – Native Hawaiian or Other Pacific Islander 5 – White 6 – Information not provided by borrower in mail, Internet, or telephone application 8 – No Co-borrower
76	CoRace5	Co-Borrower Race 5 1 – American Indian or Alaska Native 2 – Asian 3 – Black or African American 4 – Native Hawaiian or Other Pacific Islander 5 – White 6 – Information not provided by borrower in mail, Internet, or telephone application 8 – No Co-borrower
77	HOEPA	Home Ownership and Equity Protection Act (HOEPA) Status 1 – Subject to HOEPA 2- Not subject to HOEPA
78	LienStatus	Lien Status 1 – Secured by first lien 2 – Secured by a subordinate lien 3 – Not secured by a lien 4 – Not Applicable (ie purchased loans)
79	SpcHsgGoals	Special Housing Goals Loan 1 – Yes 2 – No
80	FedFinStbltyPlan	Federal Financial Stability Plan 1 – Yes 2 – No
81	AcqTyp	Acquisition Type 1 - Cash 2 - SWAP 3 - Other 4 - Credit Enhancement 5 - Bond or Debt Purchase 6 - REMIC 7 - Reinsurance 8 - Risk Sharing 9 - REIT
82	GSEREO	GSE Real Estate Owned 1 – Yes 2 – No

Field Meanings

These are the 11 banks that make up the Federal Home Loan Banks San Francisco has the most expensive loans, then Boston and Chicago follow behind.

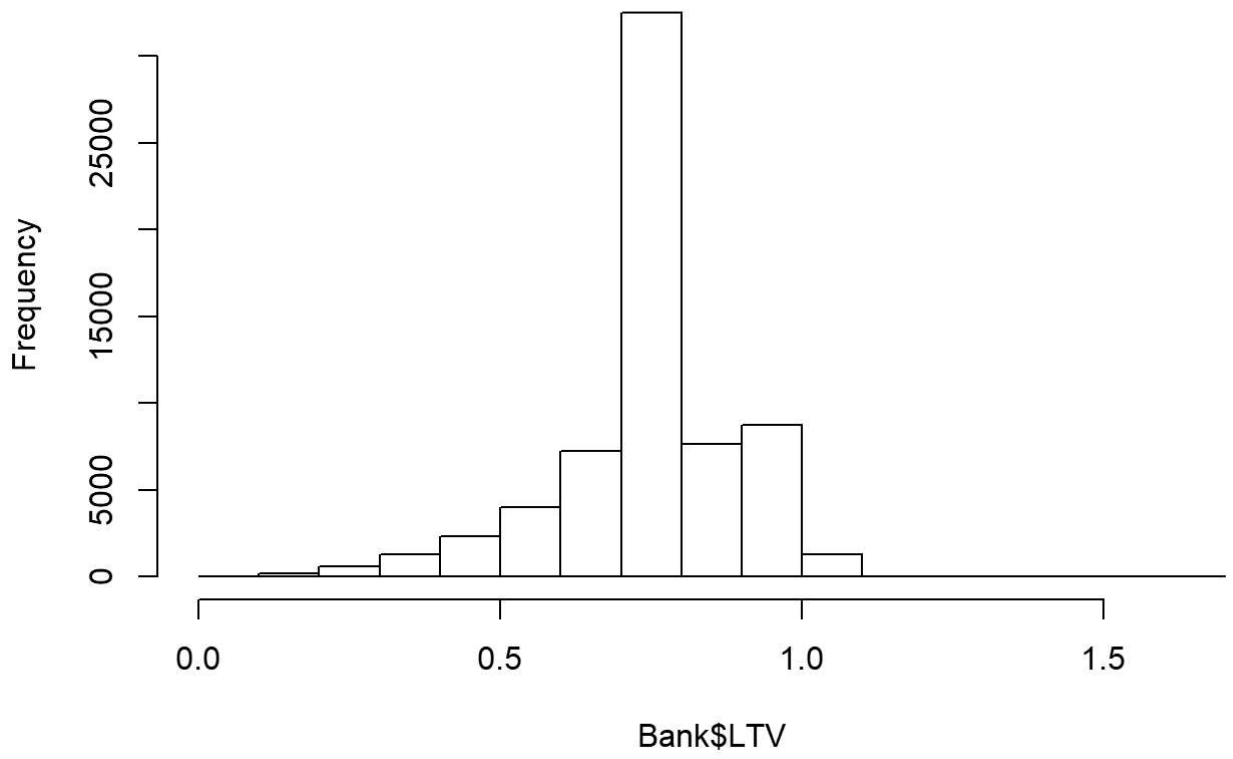
```
qplot(data=Bank, x=Bank, y=Amount)
```




```
#Total Borrowers annual income in whole dollars
#Summary shows median in $95K, max is $9.6MM

summary(Bank$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11004   63492   95000  112689  137035 9614088
```

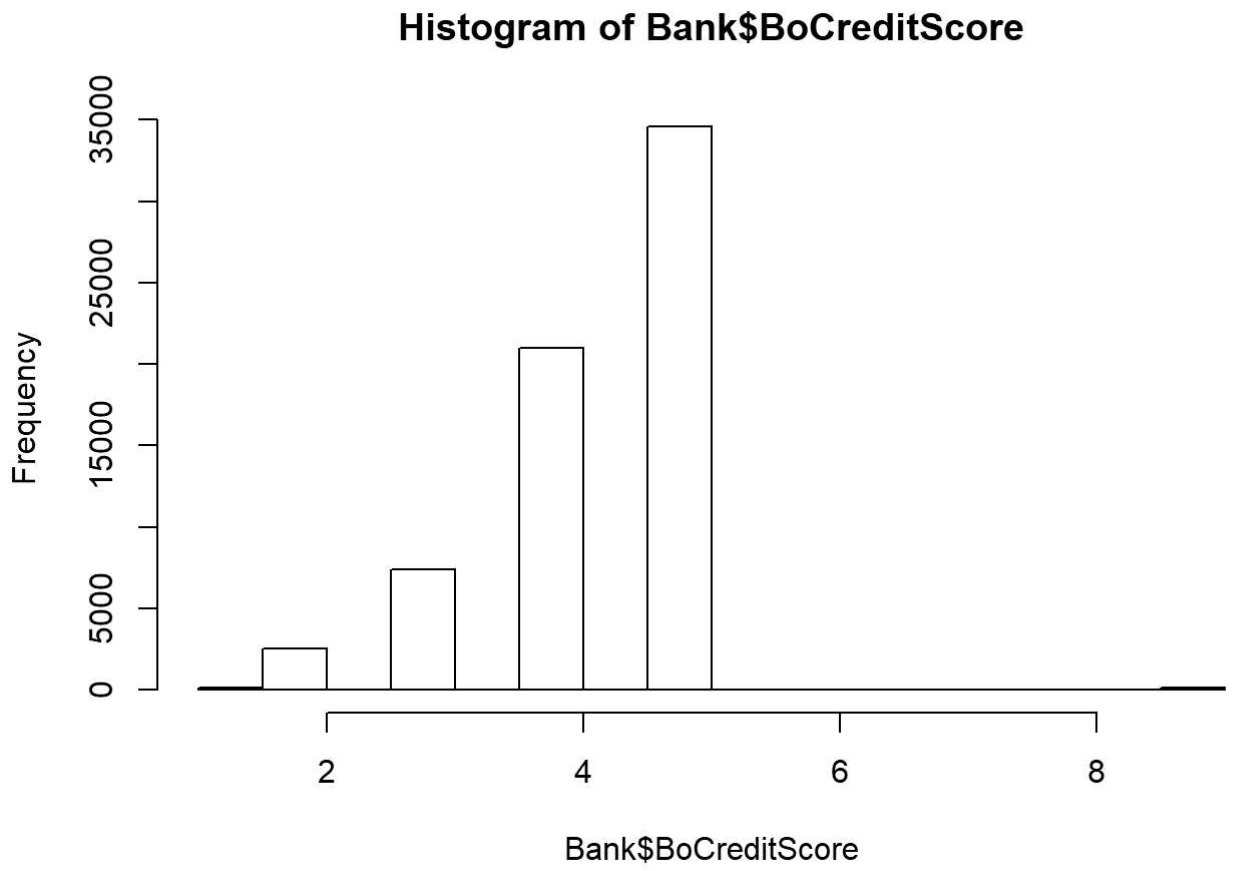
```
#Histogram of LTV(Loan to Value Ratio)
#Most LTVs fall within above 50% and below 100%, majority in 70%
hist(Bank$LTV)
```




```
summary(Bank$LTV)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.008   0.720   0.800   0.771   0.850   1.680
```

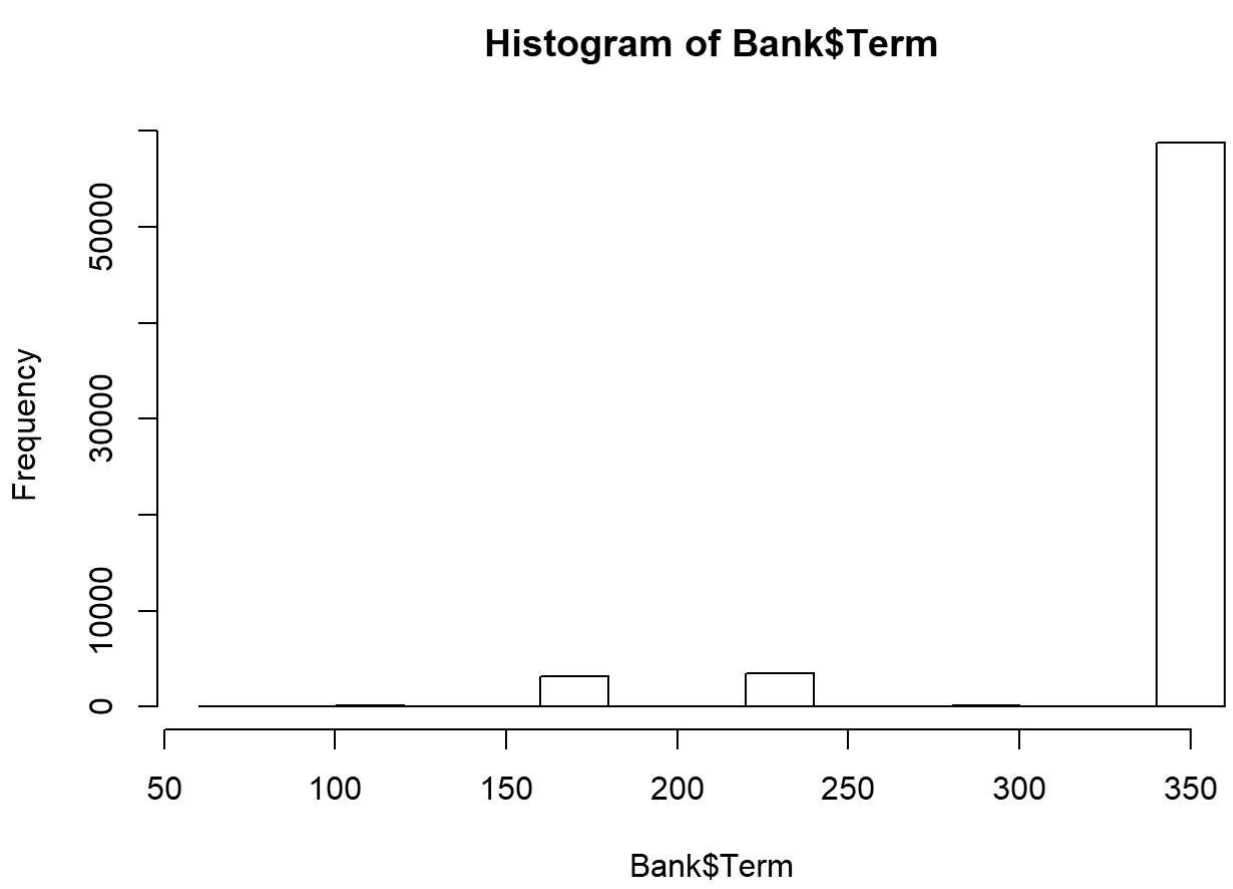
```
#Histogram of Borrowers credit score
hist(Bank$BoCreditScore)
```




```
summary(Bank$BoCreditScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   4.342   5.000   9.000
```

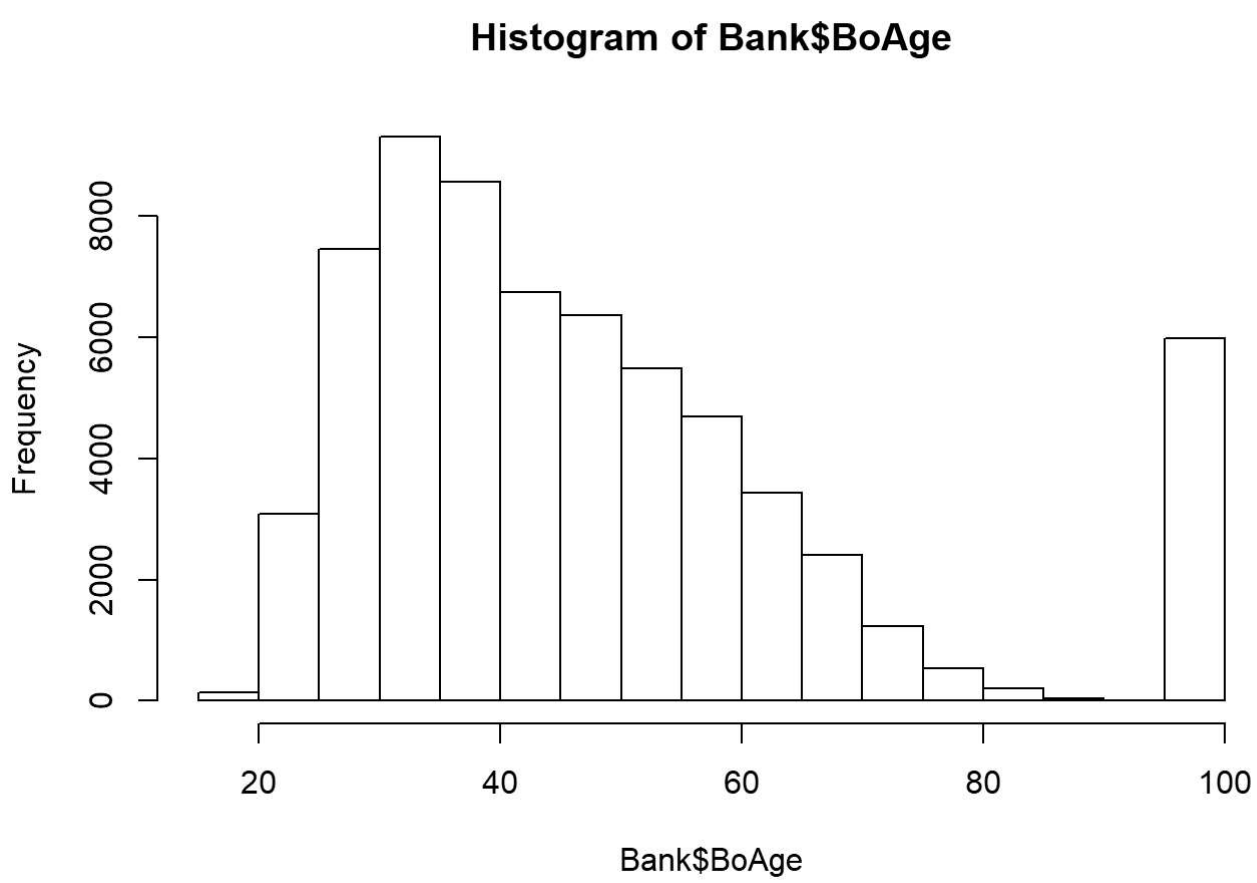
```
#Credit scores are seperated into ranges:
#1 <620
#2 620 to <660
#3 660 < 700
#4 700 < 760
#5 760 or greater
#9 Missing

#Historgram of Terms 15YR/180 20YR/240 30YR/360
#Most are clustered in these terms
hist(Bank$Term)
```




```
summary(Bank$Term)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   360.0   360.0   344.3   360.0   360.0
```

```
#Histogram of Borrowers Age, most below 40, 99 means Not Provided
hist(Bank$BoAge)
```



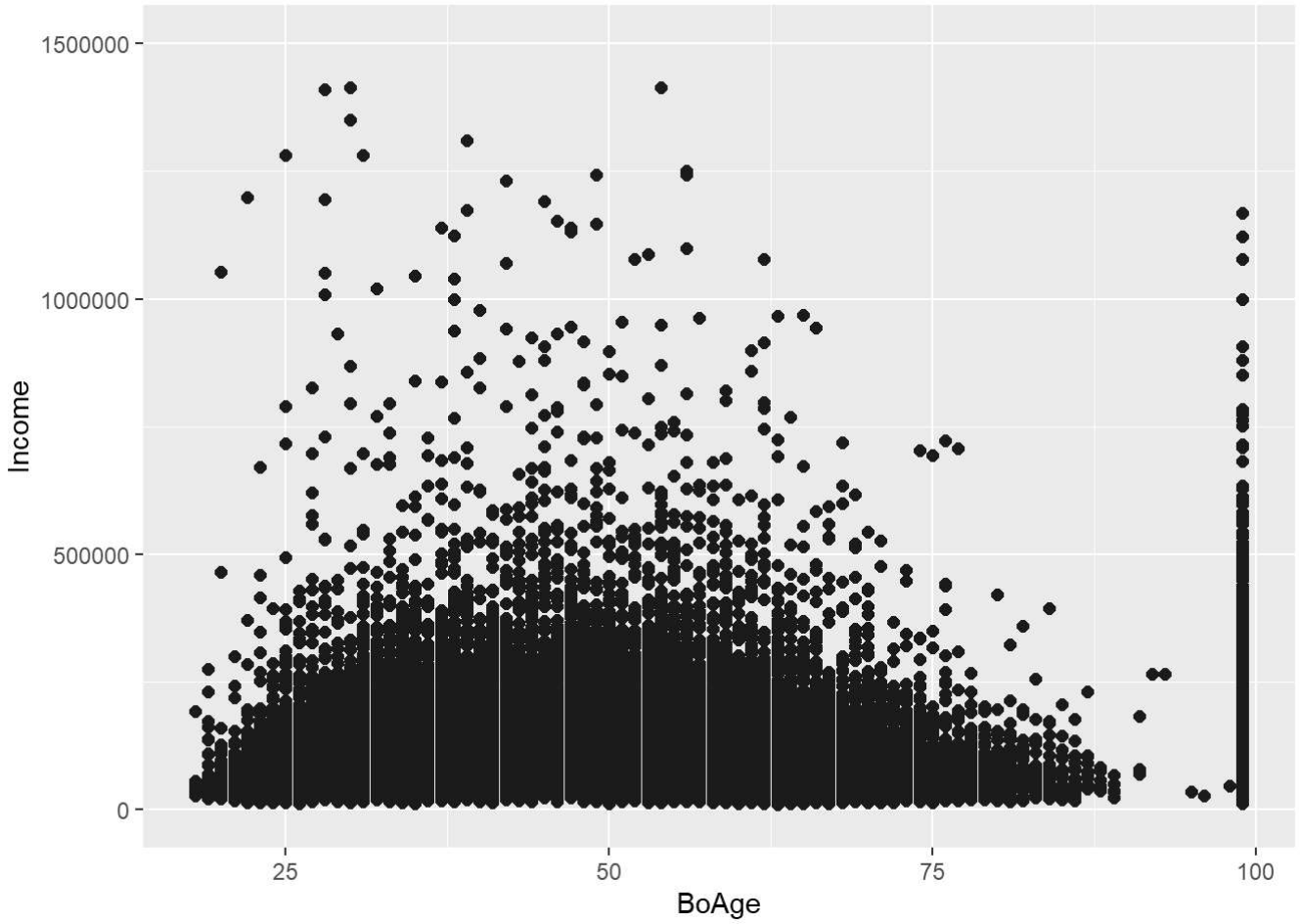

```
summary(Bank$BoAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    34.0    44.0    48.7    58.0    99.0
```

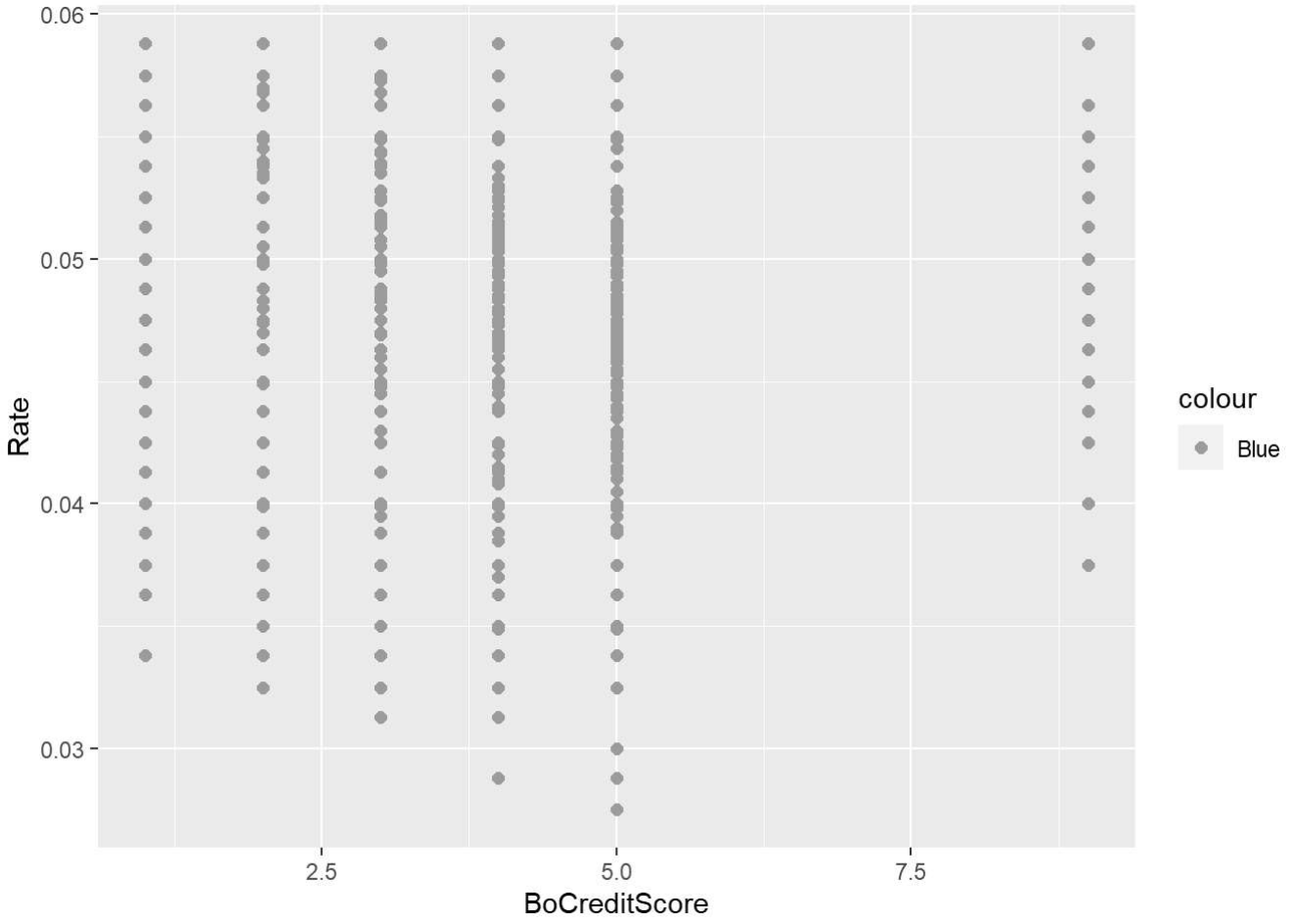
Let's take a look at Age vs Income

```
#Age vs Income
#Limited the Y range due to some larger outliers, also limited X due to 99 is all N/A
ggplot( Bank, aes(x=BoAge, y=Income)) + geom_point(size=2,color="Blue") + scale_y_continuous(lim
its = c(0, 1500000))
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```



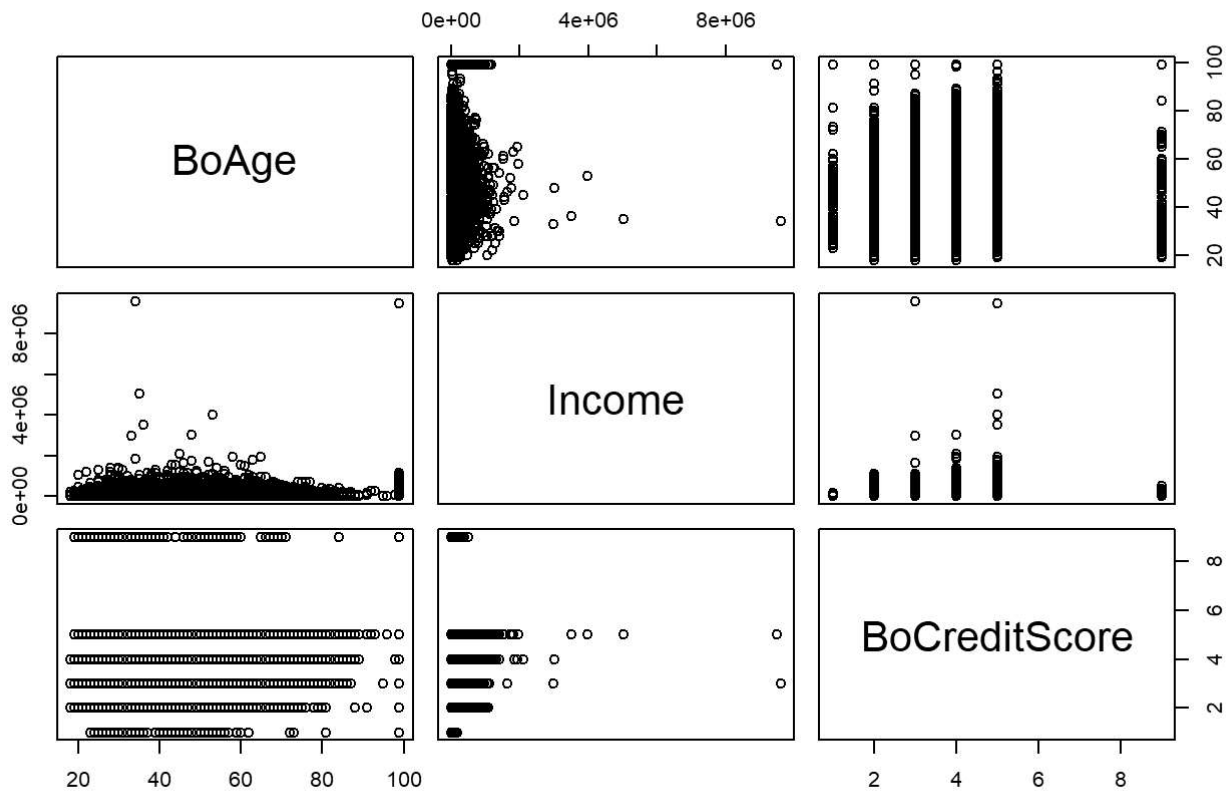

```
#Income is varied throughout age range, few outliers, but no linear increase over time. It actua
lly appears to decline after it peaks.

#Let's also try Rate x Credit score, which traditionally is correlated as Credit Score is a valu
e used in determining interest rate:
qplot(data=Bank, y=Rate, x=BoCreditScore, color= "Blue", size=I(2))
```



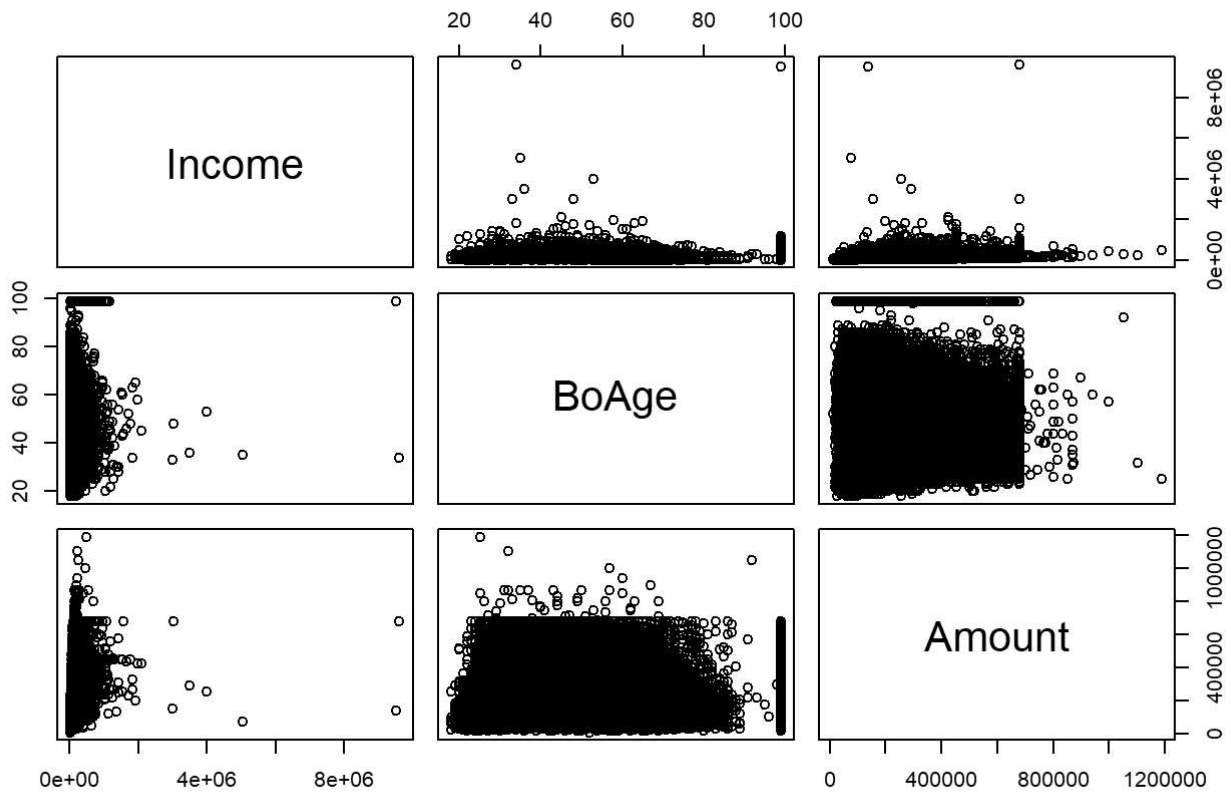

```
#1 <620
#2 620 to <660
#3 660 < 700
#4 700 < 760
#5 760 or greater
#9 Missing
#You can clearly see above as the credit score increases, the rate does get lower, although ther
e are still high rates througout all
#credit score levels.
```

Lets take a look at some different plots/charts to see if we find anything useful:

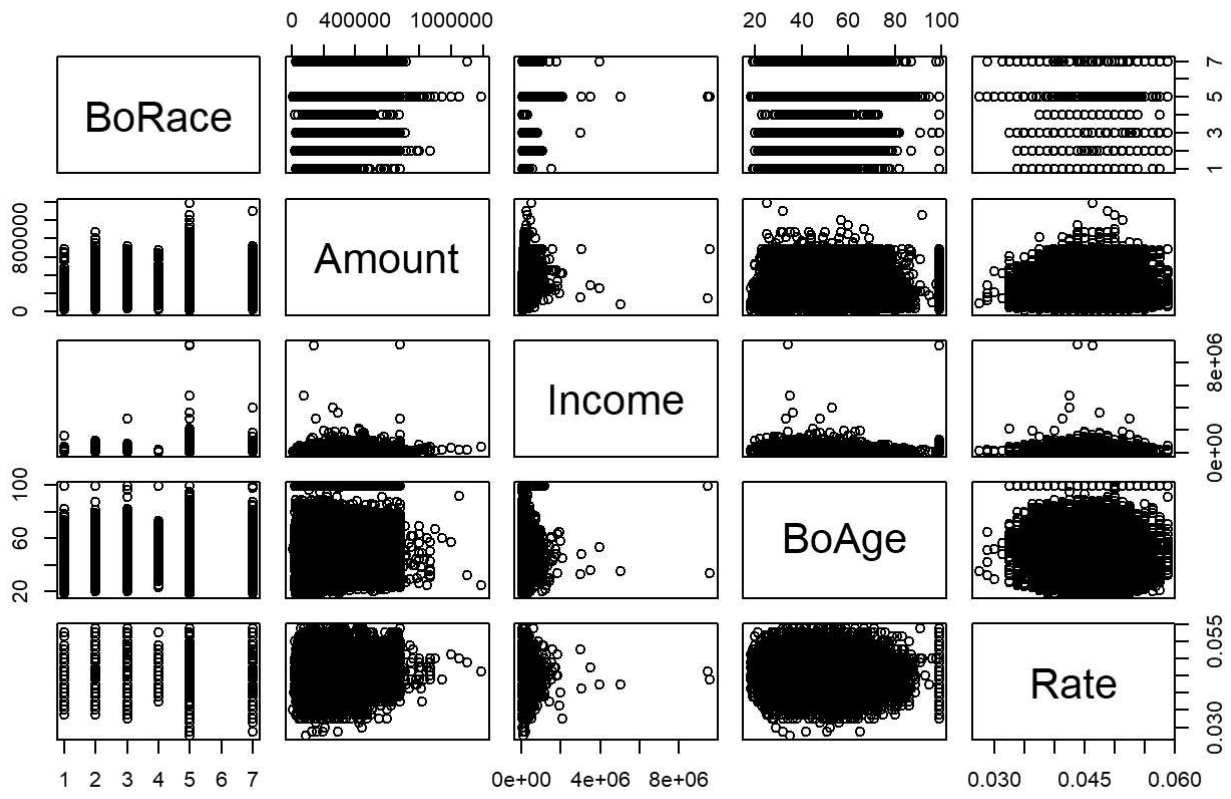
```
#Scatterplot to see the data a bit better
pairs(~BoAge+Income+BoCreditScore,data = Bank,
      main = "Scatterplot Matrix")
```



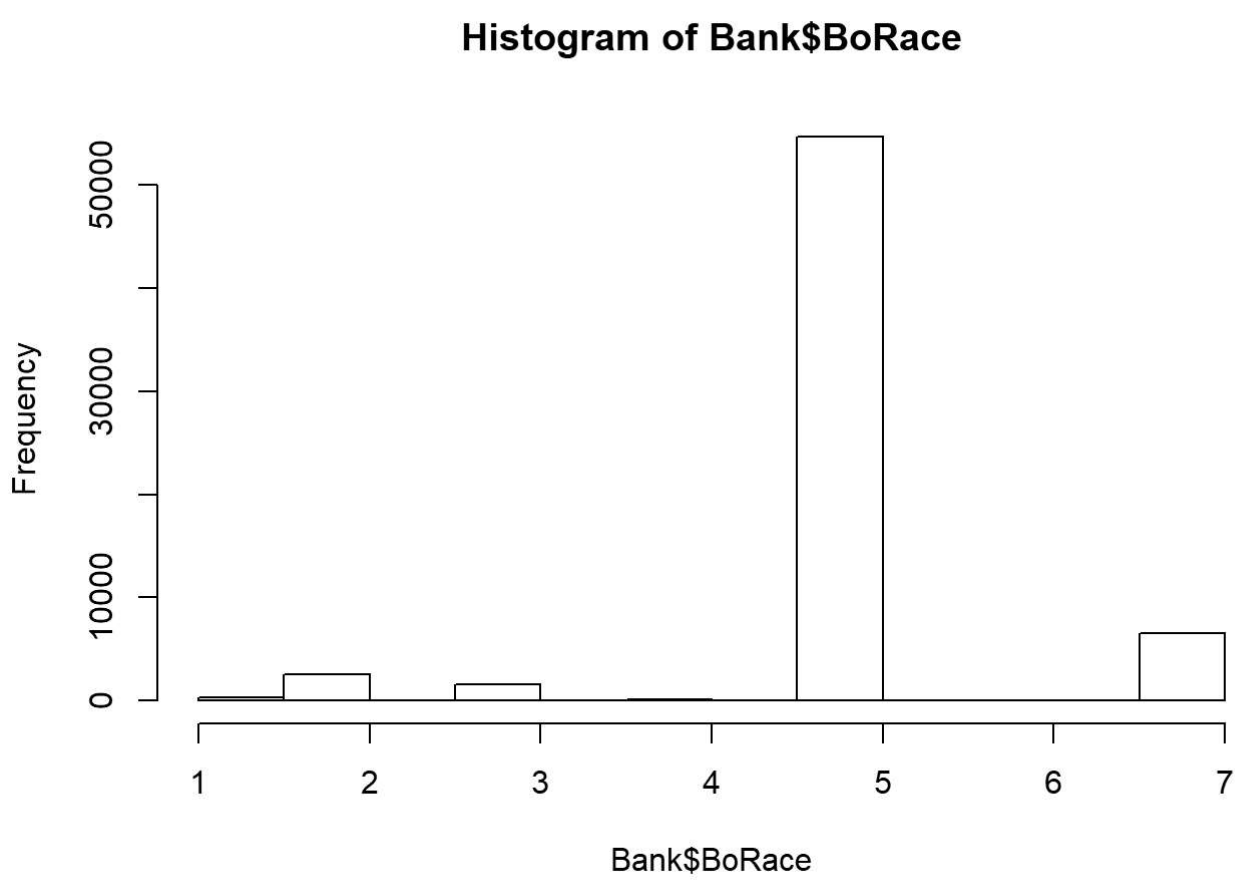

```
#Nothing really here, no linear examples noticable

#Lets try some others
pairs(~Income+BoAge+Amount,data = Bank,
      main = "Scatterplot Matrix")
```



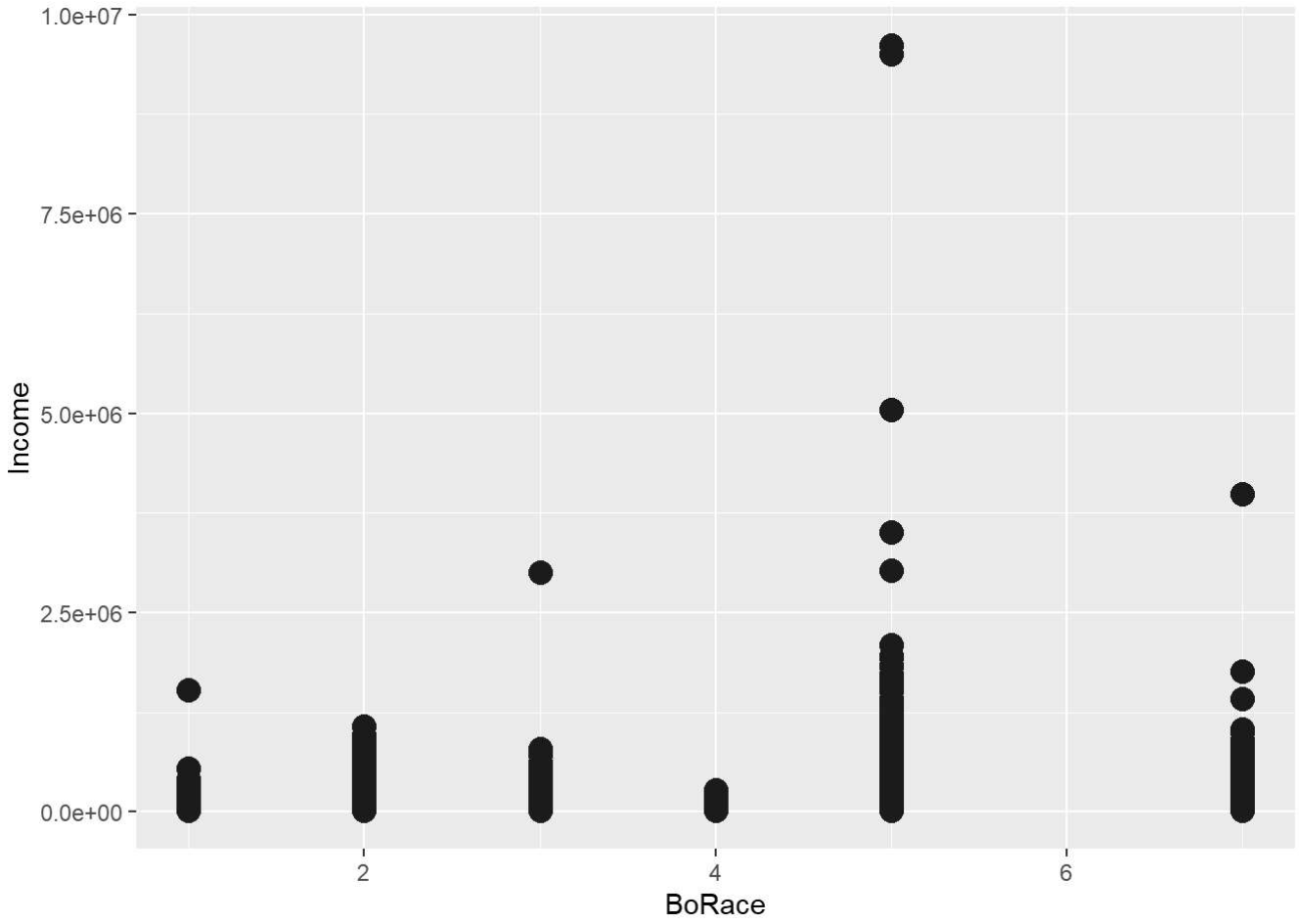

```
#Nothing useful, as the general trend seems to be that the data groupings are closely packed tog
ether and do not appear to be varied much.


#Lets try some others
pairs(~BoRace+Amount+Income+BoAge+Rate,data = Bank,
      main = "Scatterplot Matrix")
```




```
#According to histogram, white race by far has the most loans.
hist(Bank$BoRace)
```




```
#As well as income x race is clearly higher for code 5 which is equal to white
options(Scipen=999)
#Above is remove the scientific notation although it doesn't appear to work in this example but has eleswhere.
ggplot( Bank, aes(x=BoRace, y=Income)) + geom_point(size=4,color="Blue")
```




```
#1 = American Indian or Alaskan Native;
#2 = Asian;
#3 = Black or African American;
#4 = Native Hawaiian or other Pacific Islander;
#5 = White;
#7 = Information not provided
```

```
#NEXT STEP BELOW FOR TRYING TO FIND LM BETWEEN BoRace & Income, not necessarily linear though
#Checking LM for lm(y~x)
resultsRace = lm ( Bank$BoRace~Bank$Income+Bank$Amount )
resultsRace
```

```
##
## Call:
## lm(formula = Bank$BoRace ~ Bank$Income + Bank$Amount)
##
## Coefficients:
## (Intercept)  Bank$Income  Bank$Amount
##   5.049e+00    2.708e-07   -2.776e-07
```

```
summary(resultsRace)
```

```
##
## Call:
## lm(formula = Bank$BoRace ~ Bank$Income + Bank$Amount)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3912 -0.0364 -0.0195  0.0058  2.1924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.049e+00  7.826e-03 645.173  < 2e-16 ***
## Bank$Income  2.709e-07  4.010e-08   6.754 1.45e-11 ***
## Bank$Amount -2.775e-07  3.101e-08  -8.952  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9569 on 65700 degrees of freedom
## Multiple R-squared:  0.001391,   Adjusted R-squared:  0.00136
## F-statistic: 45.75 on 2 and 65700 DF,  p-value: < 2.2e-16
```

```
#P value and R squared both extremely small so no correlation
#Adding another variable below
resultsRace2 = lm ( Bank$BoRace~Bank$Income+Bank$Amount+Bank$BoGender + Bank$BoAge )
resultsRace2
```

```
##
## Call:
## lm(formula = Bank$BoRace ~ Bank$Income + Bank$Amount + Bank$BoGender +
##     Bank$BoAge)
##
## Coefficients:
##   (Intercept)    Bank$Income    Bank$Amount  Bank$BoGender     Bank$BoAge
##     4.360e+00      3.885e-07     -3.371e-07      4.595e-01      9.476e-04
```

```
summary(resultsRace2)
```

```
##
## Call:
## lm(formula = Bank$BoRace ~ Bank$Income + Bank$Amount + Bank$BoGender +
##     Bank$BoAge)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8138 -0.2730  0.1519  0.1912  2.4297
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.360e+00  1.361e-02 320.349  < 2e-16 ***
## Bank$Income    3.885e-07  3.840e-08  10.116  < 2e-16 ***
## Bank$Amount   -3.371e-07  2.965e-08 -11.366  < 2e-16 ***
## Bank$BoGender  4.595e-01  5.881e-03  78.140  < 2e-16 ***
## Bank$BoAge     9.476e-04  1.753e-04   5.406 6.47e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9144 on 65698 degrees of freedom
## Multiple R-squared:  0.0881, Adjusted R-squared:  0.08805
## F-statistic:  1587 on 4 and 65698 DF,  p-value: < 2.2e-16
```

```
#Checking LM for lm(y~x)
#Checking if LTV is related to rate or amount, most likely rate will be dependent on LTV
results2 = lm ( Bank$LTV~Bank$Rate+Bank$Amount )
results2
```

```
##
## Call:
## lm(formula = Bank$LTV ~ Bank$Rate + Bank$Amount)
##
## Coefficients:
## (Intercept)    Bank$Rate  Bank$Amount
##   5.689e-01    4.361e+00    1.490e-08
```

```
summary(results2)
```

```
##
## Call:
## lm(formula = Bank$LTV ~ Bank$Rate + Bank$Amount)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76682 -0.05455  0.02298  0.07389  0.88505
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.689e-01  7.399e-03  76.887  < 2e-16 ***
## Bank$Rate   4.361e+00  1.613e-01  27.043  < 2e-16 ***
## Bank$Amount 1.490e-08  4.347e-09   3.429 0.000607 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.147 on 65700 degrees of freedom
## Multiple R-squared:  0.01131,    Adjusted R-squared:  0.01128
## F-statistic: 375.8 on 2 and 65700 DF,  p-value: < 2.2e-16
```

```
#Above model shows very low P value on Amount, and R-Squared value, meaning LTV has very little effect and is not very accurate for Rate or amount

#Checking if LTV is related to rate and removing amount
results3 = lm ( Bank$LTV~Bank$Rate)
results3
```

```
##
## Call:
## lm(formula = Bank$LTV ~ Bank$Rate)
##
## Coefficients:
## (Intercept)    Bank$Rate
##      0.5714       4.3824
```

```
summary(results3)
```

```
##
## Call:
## lm(formula = Bank$LTV ~ Bank$Rate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76632 -0.05432  0.02568  0.07568  0.88377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.571408   0.007362   77.62   <2e-16 ***
## Bank$Rate   4.382440   0.161139   27.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.147 on 65701 degrees of freedom
## Multiple R-squared:  0.01113,    Adjusted R-squared:  0.01112
## F-statistic: 739.7 on 1 and 65701 DF,  p-value: < 2.2e-16
```

```
#Doesn't appear to be depedent on rate, which is interesting as you would think they are depende
nt on one another?


#Also checking to see if Credit score shows a linear relationship to Rate
lrRate<- lm(Bank$BoCreditScore~Bank$Rate)
lrRate
```

```
##
## Call:
## lm(formula = Bank$BoCreditScore ~ Bank$Rate)
##
## Coefficients:
## (Intercept)    Bank$Rate
##       6.299      -42.962
```

```
summary(lrRate)
```

```
##
## Call:
## lm(formula = Bank$BoCreditScore ~ Bank$Rate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8469 -0.3657  0.3679  0.6343  5.2272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.29898    0.04237  148.67   <2e-16 ***
## Bank$Rate   -42.96215    0.92741  -46.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8463 on 65701 degrees of freedom
## Multiple R-squared:  0.03163,    Adjusted R-squared:  0.03161
## F-statistic:  2146 on 1 and 65701 DF,  p-value: < 2.2e-16
```

```
#Although there is some correlation, the R squared and P value (0.03163 and .0000) all show almo
st no relationship in this fashion.
```

In summary I have looked into many of the different fields provided by this dataset and although you can see some difference in data there doesn't seem to be any major correlations in the data. I believe part of the lack of field correlations is that all of the major values are somewhat grouped together. What I mean by that is that many of the lending criteria are closely related. For example when looking at income, most is grouped together over a certain threshold and follows a partial bell curve based on age, but to find a linear correlation to this I was unsuccessful. The same follows for interest rate and loan amounts, as almost all stay in a limited range from approximately 3% and above to less than 7%. Loan amounts often have to follow conforming limits (specific dollar amounts per county, debt to income ratios, etc.), especially within these government funded loans, therefore there were not many definitive patterns as most of the data was grouped within a specific window but predictability was almost non existent within the examples I provided. I spent quite a bit of time getting this far, and in retrospect I may have found a better dataset or spent more time on mastering modifying fields. For example, I would have liked to change a field such as race to classify white vs non-white to complete a logistic regression analysis on this specific question regarding the amount of loans found. There seemed to be a relationship in that field. Perhaps in the future I can spend more time on this analysis but unfortunately I ran out of time with other than some insightful graphs but what appears to be a definitive answer that there is no linear relationships between the data fields present based on the chart imagery and the values found in the R squared and P data.