Wine Analysis draft

```
# read in ggplot2 and load red wine dataset
library(ggplot2)
redwine <- read.csv('wineQualityReds.csv')

head(redwine)
```

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

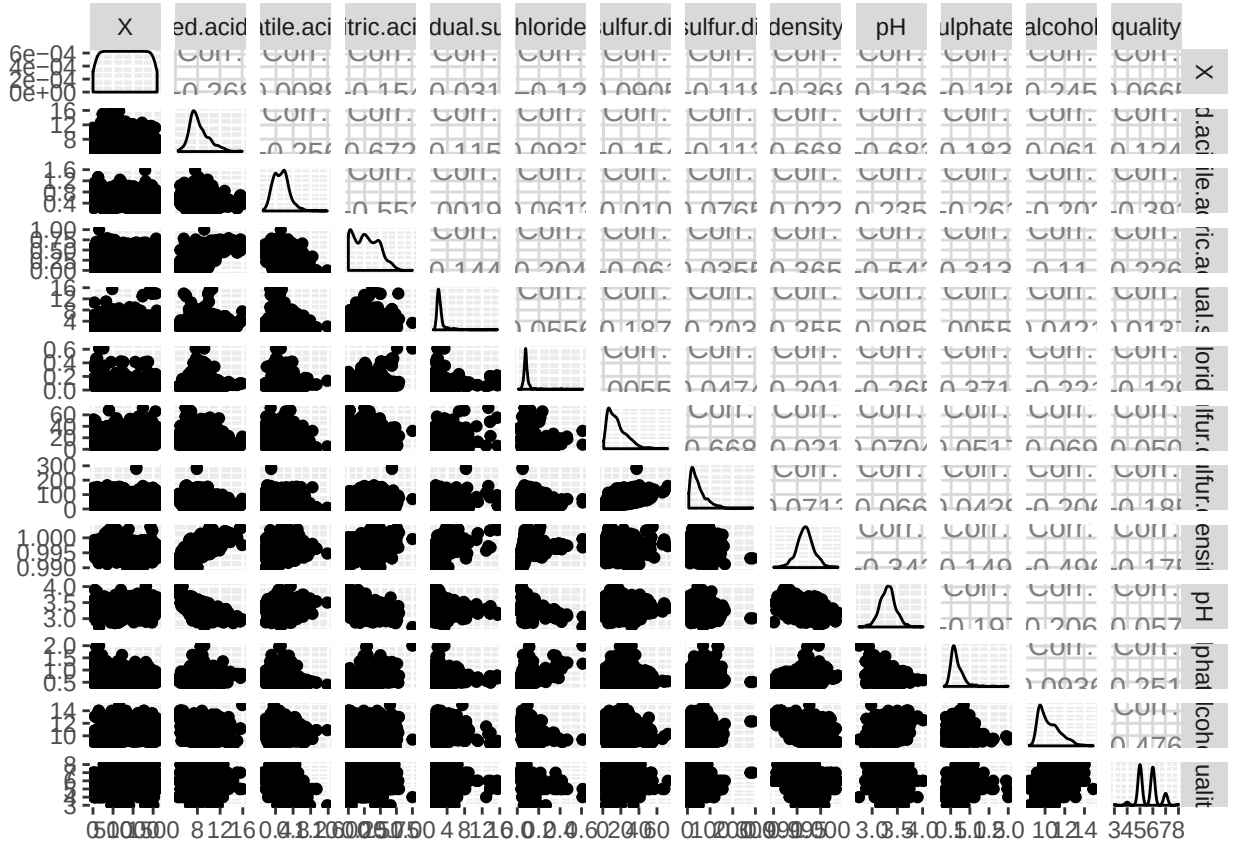
```
names(redwine)
```

```
##  [1] "X"                   "fixed.acidity"       "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"      "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"           "alcohol"
## [13] "quality"
```

```
str(redwine)
```

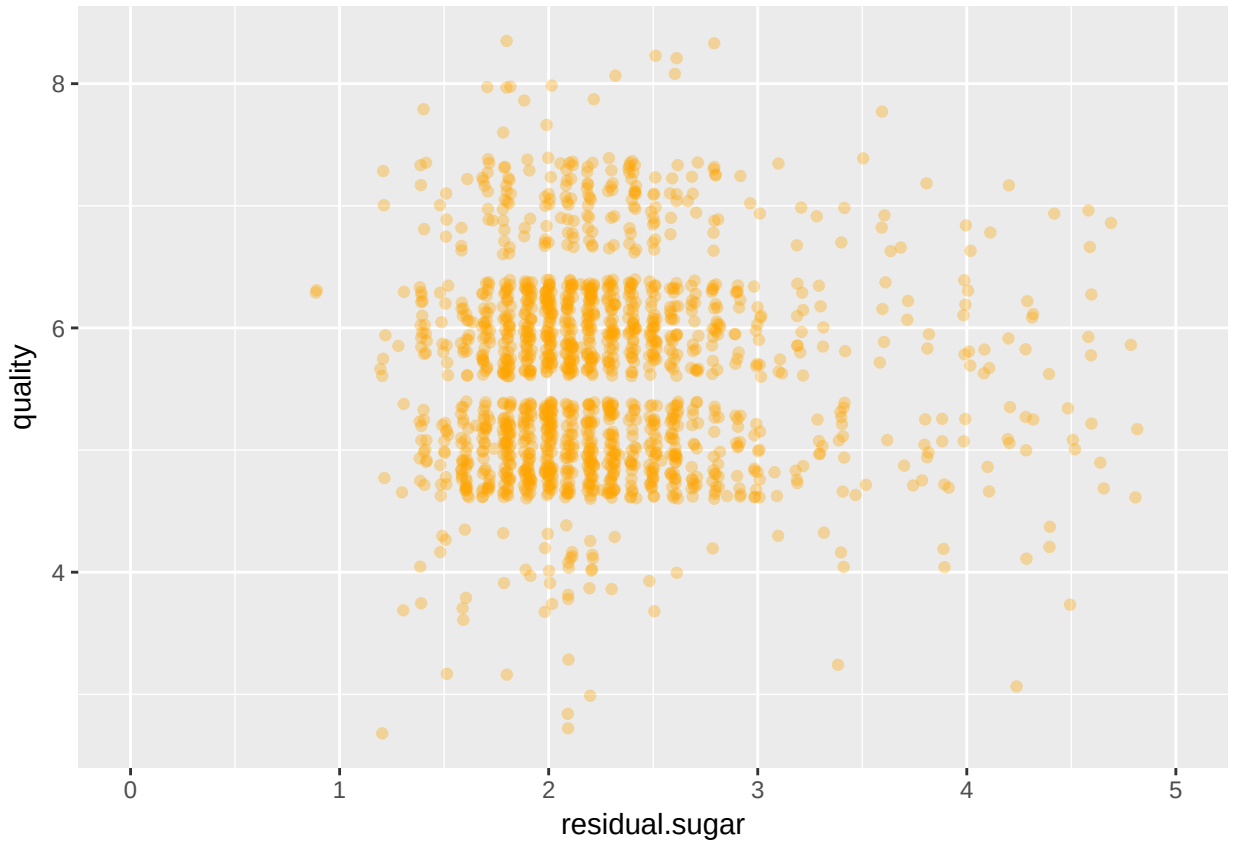
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
# doing a plot with ggpairs
library(GGally)
ggpairs(redwine)
```




```
# looking at if alcohol and quality is a correlation at all.

ggplot(data = redwine, aes(x = residual.sugar, y = quality)) +
  geom_jitter(alpha = 7/20, color = 'orange') +
  scale_x_continuous(limits = c(0,5))
```

```
## Warning: Removed 85 rows containing missing values (geom_point).
```




```
loandata <- read.csv('prosperLoanData.csv')

#print loan data
str(loandata)
```

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                          : Factor w/ 113066 levels "00003546482094282EF90E5",..: 7180 7
##  $ ListingNumber                       : int  193129 1209647 81716 658116 909464 1074836 750899 768193
##  $ ListingCreationDate                 : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..: 1
##  $ CreditGrade                         : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1 1 1 .
##  $ Term                                : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                          : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4
##  $ ClosedDate                          : Factor w/ 2803 levels "","2005-11-25 00:00:00",..: 1138 1 120
##  $ BorrowerAPR                         : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate                        : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield                         : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedEffectiveYield             : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                       : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                     : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.             : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.               : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 .
##  $ ProsperScore                        : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.           : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                       : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18
##  $ Occupation                          : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43
##  $ EmploymentStatus                    : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ EmploymentStatusDuration            : int  2 44 NA 113 44 82 172 103 269 269 ...
```

```
##  $ IsBorrowerHomeowner               : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CurrentlyInGroup                  : Factor w/ 2 levels "False","True": 2 1 2 1 1 1 1 1 1 1 ...
##  $ GroupKey                          : Factor w/ 707 levels "","00343376901312423168731",..: 1 1 335
##  $ DateCreditPulled                  : Factor w/ 112992 levels "2005-11-09 00:30:04.487000000",..: 1
##  $ CreditScoreRangeLower             : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper             : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ FirstRecordedCreditLine           : Factor w/ 11586 levels "","1947-08-24 00:00:00",..: 8639 661
##  $ CurrentCreditLines                : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                   : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years        : int  12 29 3 29 49 49 20 10 32 32 ...
##  $ OpenRevolvingAccounts             : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ OpenRevolvingMonthlyPayment       : num  24 389 0 115 220 1410 214 101 219 219 ...
##  $ InquiriesLast6Months              : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                    : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies              : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                  : num  472 0 NA 10056 0 ...
##  $ DelinquenciesLast7Years           : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years          : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months         : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance            : num  0 3989 NA 1444 6193 ...
##  $ BankcardUtilization               : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit           : num  1500 10266 NA 30754 695 ...
##  $ TotalTrades                       : num  11 29 NA 26 39 47 16 10 29 29 ...
##  $ TradesNeverDelinquent..percentage. : num  0.81 1 NA 0.76 0.95 1 0.68 0.8 1 1 ...
##  $ TradesOpenedLast6Months           : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                 : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeRange                       : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4
##  $ IncomeVerifiable                  : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
##  $ StatedMonthlyIncome               : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                           : Factor w/ 113066 levels "00003683605746079487FF7",..: 100337
##  $ TotalProsperLoans                 : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ TotalProsperPaymentsBilled        : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ OnTimeProsperPayments             : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPaymentsOneMonthPlusLate   : int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPrincipalBorrowed          : num  NA NA NA NA 11000 NA NA NA NA NA ...
##  $ ProsperPrincipalOutstanding       : num  NA NA NA NA 9948 ...
##  $ ScorexChangeAtTimeOfListing       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination        : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                        : int  19141 134815 6466 77296 102670 123257 88353 90051 121268
##  $ LoanOriginalAmount                : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000
##  $ LoanOriginationDate               : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 426 1866 260
##  $ LoanOriginationQuarter            : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32 24 33 1
##  $ MemberKey                         : Factor w/ 90831 levels "00003397697413387CAF966",..: 11071 10
##  $ MonthlyLoanPayment                : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments               : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments      : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                    : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss               : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ LP_NonPrincipalRecoverypayments    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                          : int  258 1 41 158 20 1 1 1 1 1 ...
```

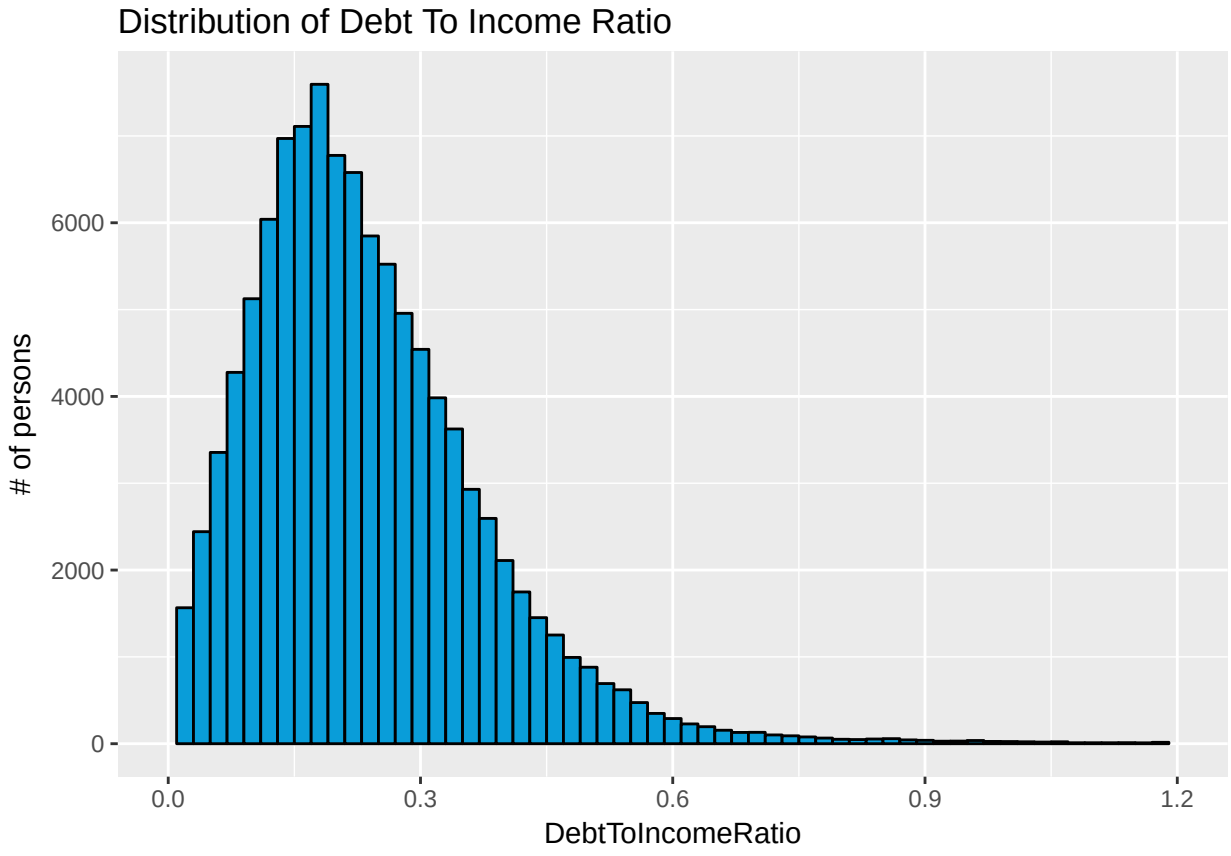
```
ggsave('name.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
## Warning: Removed 84 rows containing missing values (geom_point).
```

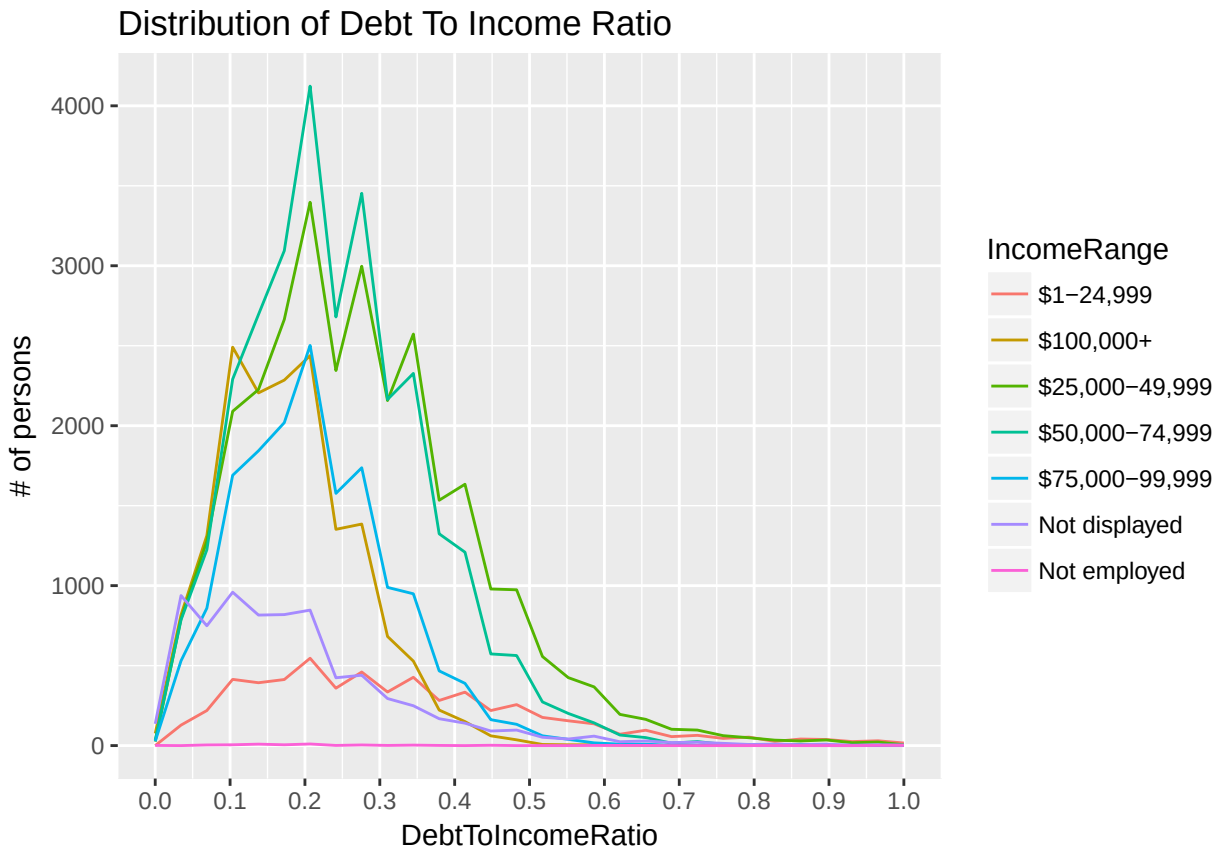
```
#Below shows the distribution of the debt to income ratio of the dataset.

ggplot(data = loandata, aes(DebtToIncomeRatio)) +
  geom_histogram(binwidth = 0.02, color = 'black', fill = '#099DD9') +
  scale_x_continuous(breaks = seq(0,1.2,.3), limits = c(0,1.2)) +
  ggtitle('Distribution of Debt To Income Ratio') +
  ylab('# of persons')
```

```
## Warning: Removed 9203 rows containing non-finite values (stat_bin).
```




```
ggplot(data = loandata, aes(DebtToIncomeRatio, colour = IncomeRange)) +
  geom_freqpoly(na.rm = TRUE) +
  scale_x_continuous(breaks = seq(0,1,.1), limits = c(0,1)) +
  ggtitle('Distribution of Debt To Income Ratio') +
  ylab('# of persons')
```

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



Is there any correlation between a person's occupation and how much he is in debt?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

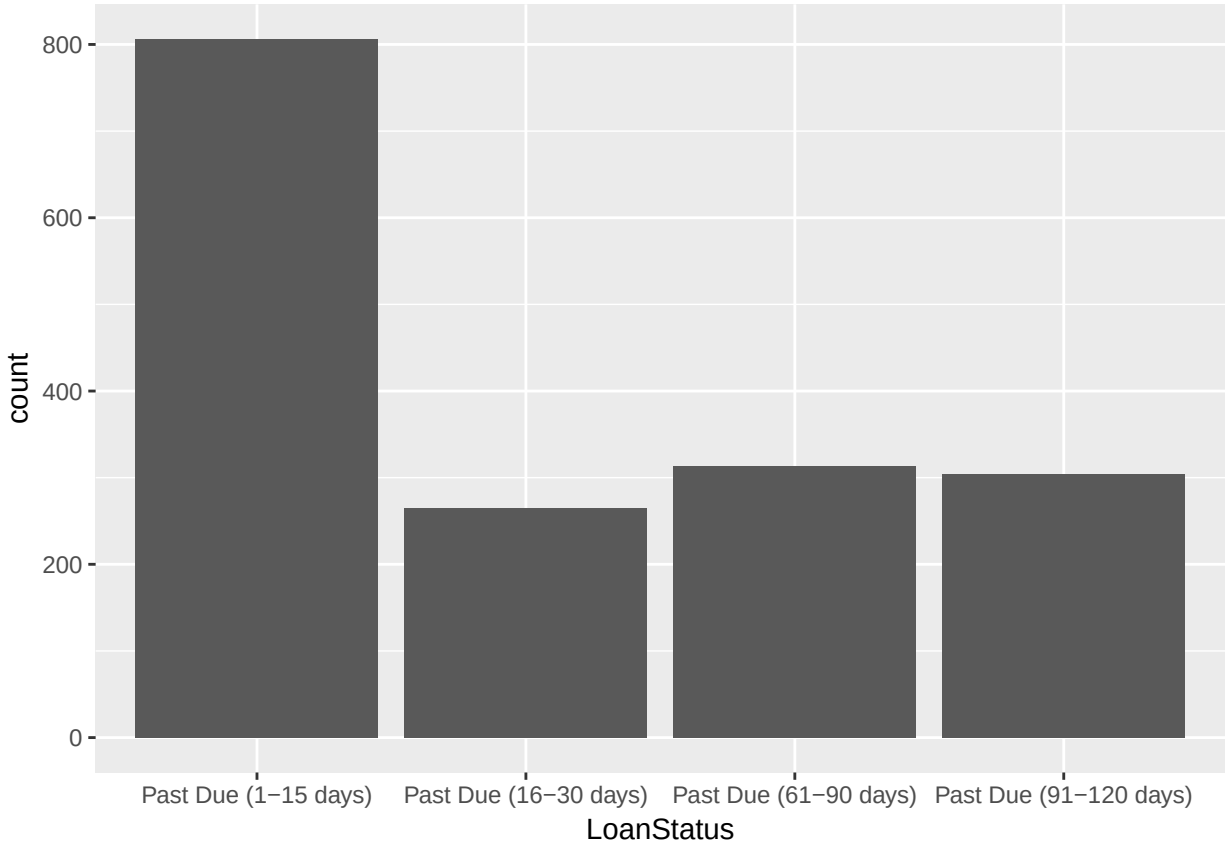
```
## The following object is masked from 'package:GGally':
##
##     nasa
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ggplot(data = subset(loandata, LoanStatus == "Past Due (16-30 days)" |
                       LoanStatus == "Past Due (>120)" |
                       LoanStatus == "Past Due (1-15 days)" |
                       LoanStatus == "Past Due (91-120 days)" |
                       LoanStatus == "Past Due (61-90 days)"),
       aes(LoanStatus)) +
  geom_histogram(stat = 'count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



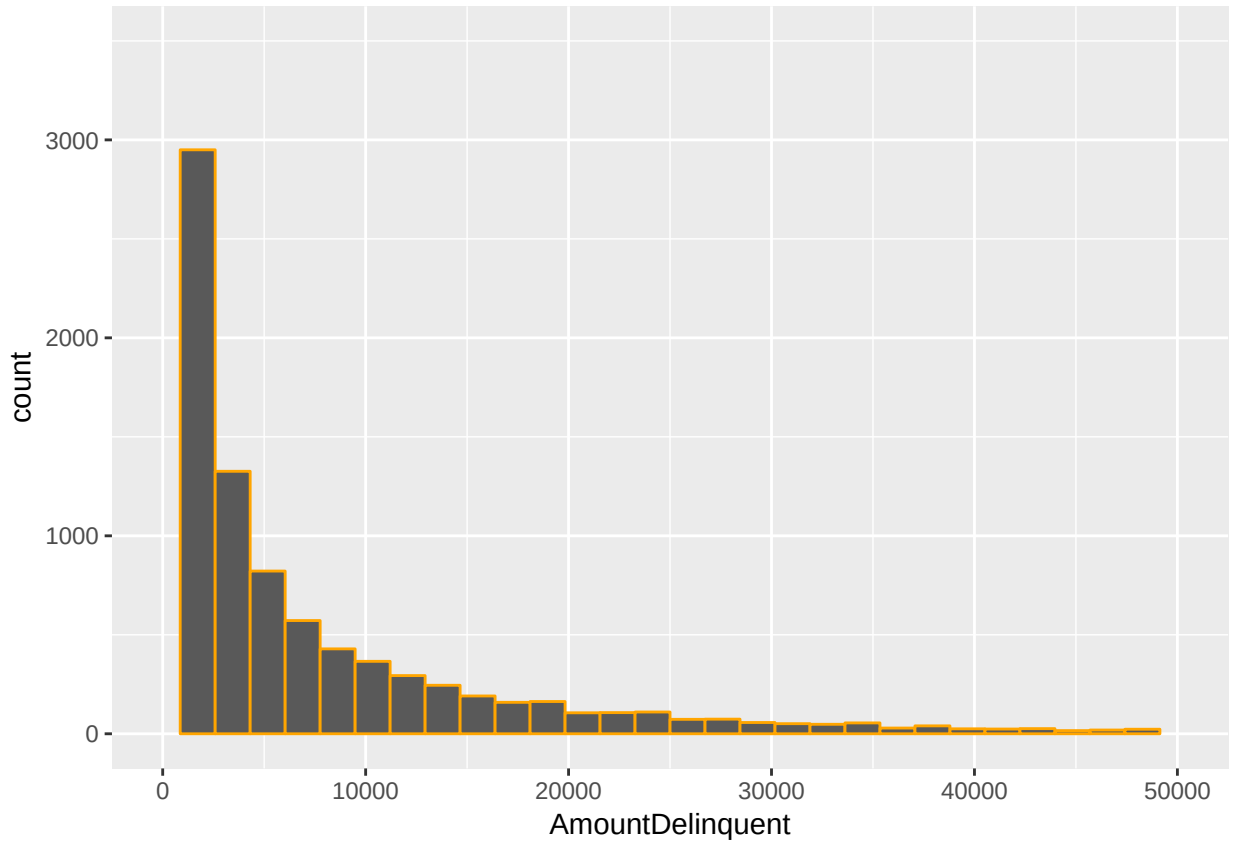
Amount Delinquent

```
ggplot(data = loandata, aes(AmountDelinquent)) +
  geom_histogram(color = 'orange') +
  scale_x_continuous(limits = c(0,50000)) +
  scale_y_continuous(limits = c(0,3500))
```

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

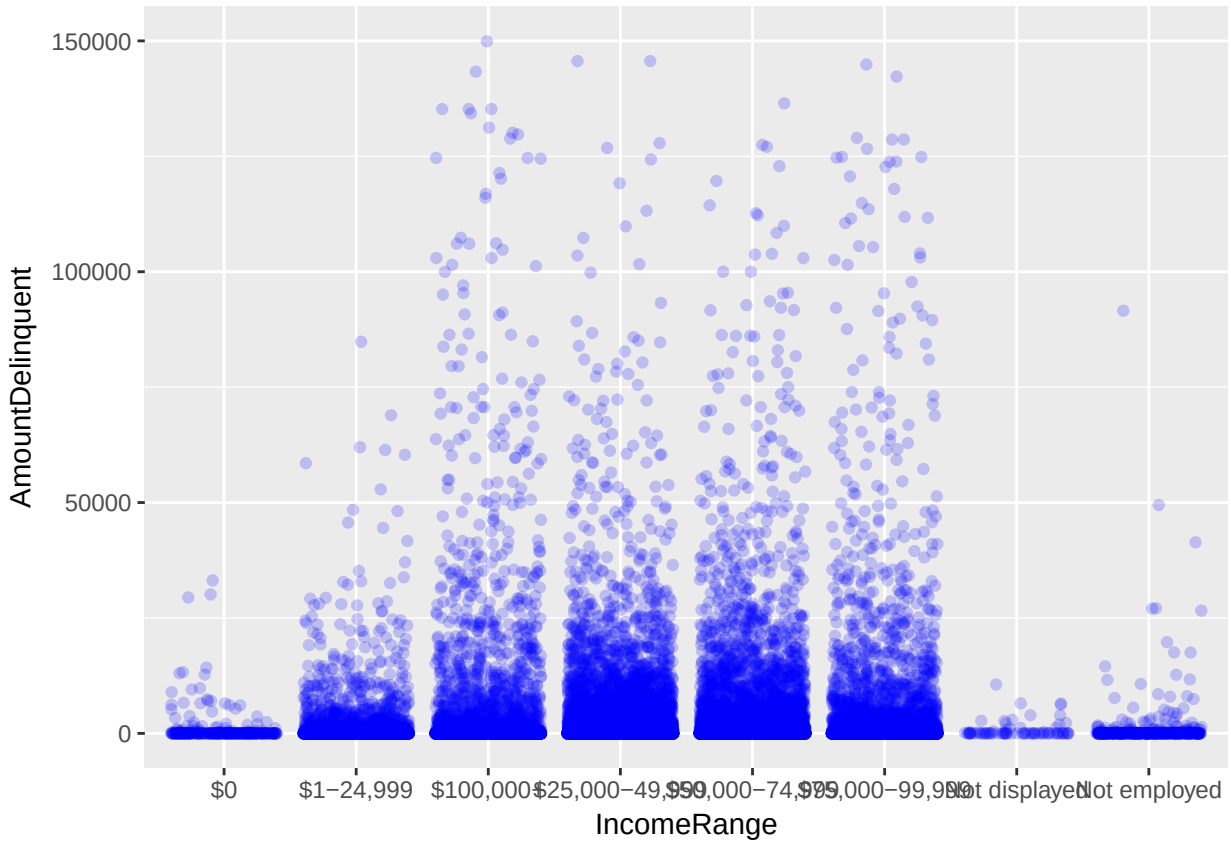
```
## Warning: Removed 7998 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```




```
#income range vs delinquent amount

ggplot(data = loandata, aes(x = IncomeRange, y = AmountDelinquent, color = Occupation)) +
  geom_jitter(alpha = 1/5, color = 'blue') +
  scale_y_continuous(limits = c(0,150000))
```

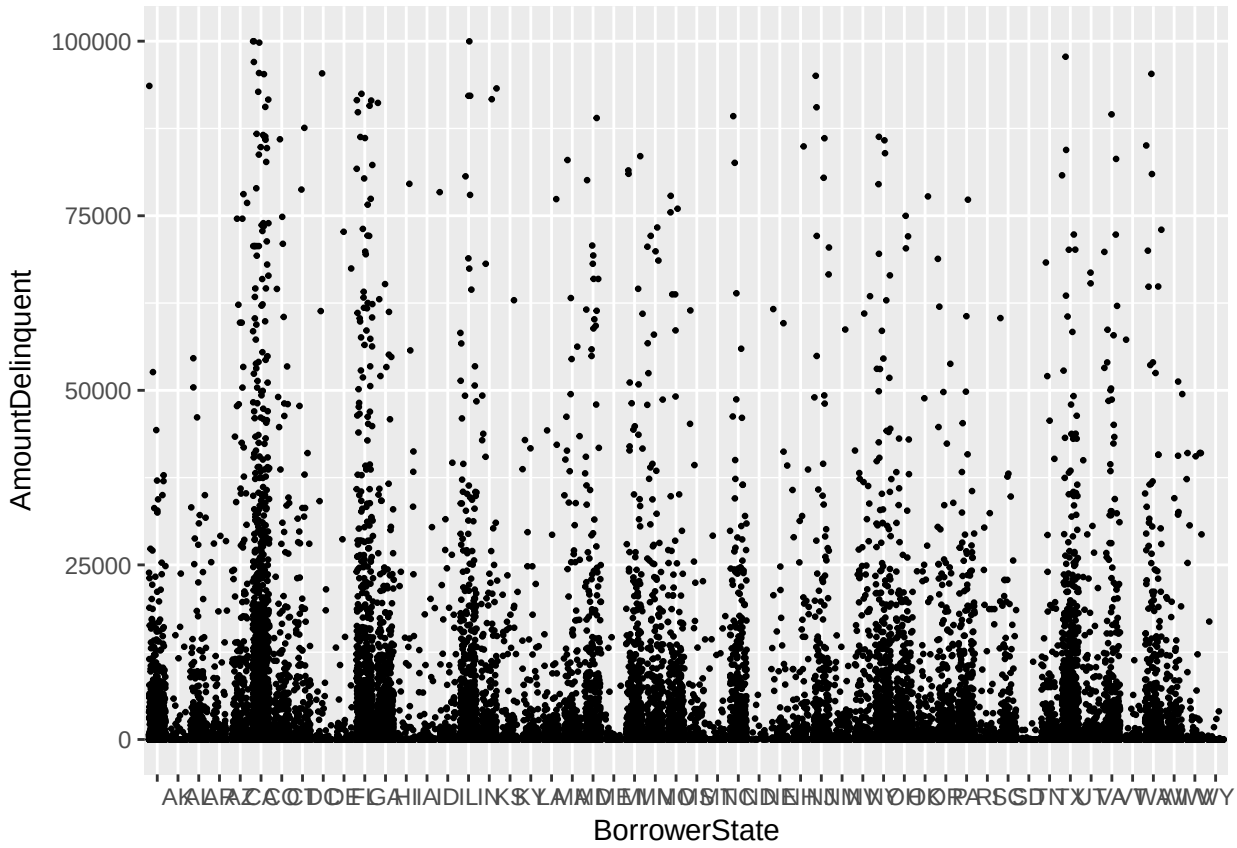
```
## Warning: Removed 52510 rows containing missing values (geom_point).
```



Do people who have higher incomes owe more? Do persons with higher incomes generally owe less because they can afford to?

```
#here is the data for places where persons live vs being in dept. Next i should look at the proportion
ggplot(data = loandata, aes(x = BorrowerState, y = AmountDelinquent)) +
  geom_jitter(size = .5) +
  scale_y_continuous(limits = c(0,100000))
```

```
## Warning: Removed 52450 rows containing missing values (geom_point).
```



As we can see above California and Florida are owing some serious money when it comes to loans