# R Project Report ： ProsperLoan公司运营及风险分析

#### v2 update note

非常感谢老师的建议,由于加了1周的其他课程测试任务,才提交变更.对于所需修改的内容,将会在本部分逐条说明,其中提出修改的图形的代码修改行也会在这里标明,辛苦老师再次审阅:

1. 代码长度问题，已经明白了可以在，处断开代码。现在是把Rstudio中左侧的长度做限制，大概要写到头了就考虑换行。请问下，有没有在Rstudio中能不能显示80字符的一条竖线（虚线）作为提示？我找了半天也没找到。
2. 补充项目信息集的描述：本次项目选择的数据集是ProsperLoan公司的数据。

* 数据量80+M，参数有81个，是项目中最大的数据集。
* 选择数据集的原因是想探索一下带宽的审批和信用的评级的数据，发现之间的联系。
* 看看Prosper公司在运营中是怎么将每一个贷款人的数据数字化并做出判断的。

1. 这也是数据的一个重要价值，不由人参与即可对客户进行细分（信用不同）并对不同细分进行针对服务。
2. 单变量的分析改进：

* 增加了新的分析指标Ratio/Income
* 所有指标的分析进行了更新
* 请参见PartA
* 所有项增加了summary的输出

1. 增加所有图形Title：已经完成。
2. 感谢关于排除一个分类的展示的反馈，已经完成。更新在绘图3中已经排除Not Displayed分类。
3. Occupation那张图重新找了对应关系，使用jitter和alpha后，比较容易看出各个职业的收入分配（图没出来之前没有想到是这样的）
4. 为未来工作提供了至少一个提议或问题：

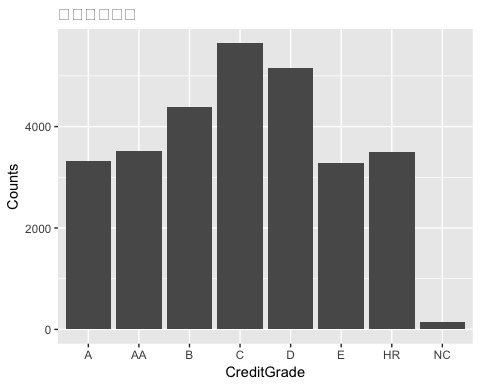
* 根据分析提出以下几个建议：
* a，在职业分类中去掉default的选择1分类，并做一些分类的调整。目前情况导致这个数据不能提供有效数据。
* b，检查0收入和not employ两个数据维度的对应情况，前者很多，后者很少，需要分析为什么
* c，从目前分类来看高等级的信用和低等级的信用都给公司带来了很多收益。但是从业务区分来讲，如果是双峰可能会更加适合业务的开展（不同侧重点），此处建议：
  + 1, 调整信用计算和其他数据的关系（比如增加带宽次数为正向指标）
  + 2，认真分析0收入和低收入的特点，从借款额初步看，也可以是比较优势的客户，需要认真分析（甚至拆分产品）
  + 3，丢弃带宽与收入比例或者调整。目前状况下此数据有很大误导。

========================================================

## (v2-updated)Part1A - 单变量绘图选择

#### CreditGrade

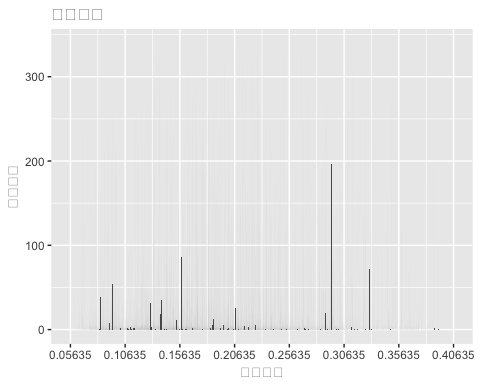
分析发现: 1. 信用评价呈现的分布比较平均 2. 众数为信用评分C级 3. 其中HR和NC为Posper公司内部分级，通用信用指标中无此分类。



## A AA B C D E HR NC NA's   
## 3315 3509 4389 5649 5153 3289 3508 141 84984

#### BorrowerAPR

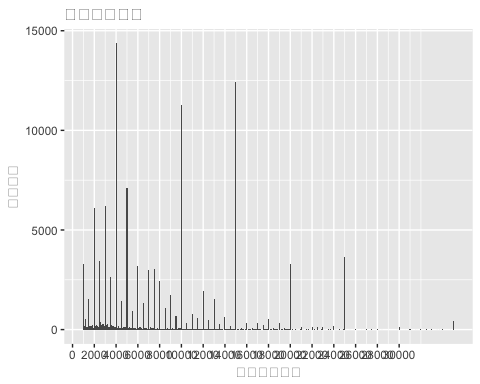
分析发现： 1. 呈现的状态为离散正态分布。 1. 中位数为0.20976。 1. 从图中可以看出：贷款利率相差很大，主要有几个档次为主要选择。



## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00653 0.15629 0.20976 0.21883 0.28381 0.51229 25

#### LoanOriginalAmount

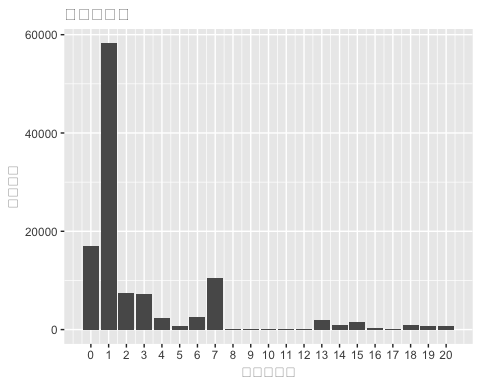
分析发现： 1. 大部分贷款低于25000美元。 1. 数量最多的发生在3000 - 4000美元。 1. 当贷款额超过10k后，每5k是一个较大聚合（10k、15k、20k）说明在可能这些贷款额是一个评估的等级。 1. 在5k以前贷款的额度比较多，在5k之后，增长是0.5k，期间的数值基本没有了。



## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00653 0.15629 0.20976 0.21883 0.28381 0.51229 25

#### ListingCategory

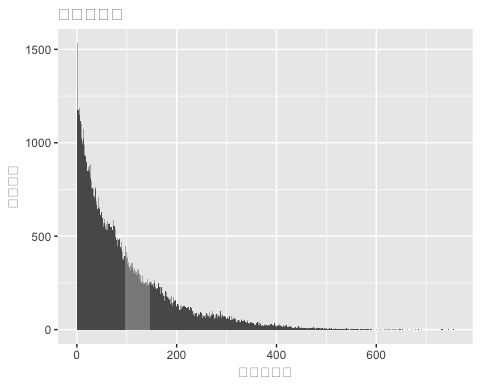
分析发现： 1. 1类为最多分类，且数量巨大。为Debt Consolidation 此分类可能是默认分类。从字面解释没有分类意义。 1. 20种分类中数量相差非常多，可能需要调整。



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 1.000 2.774 3.000 20.000

#### EmploymentStatusDuration

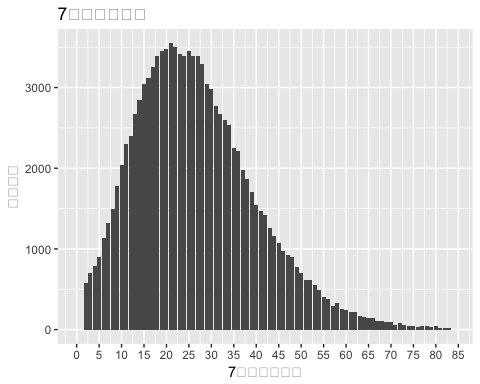
分析发现： 1. 展现了比较规范的偏斜分布数据。 2. 0的值很多，开始考虑可能和默认值有关。 1. 但如果我是贷款人，填写受雇用时长的时候，应该会非常谨慎并尽量表现得积极。所以这个趋势应该就是贷款公司客户的特点的可能性较多。



## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 26.00 67.00 96.07 137.00 755.00 7625

#### TotalCreditLinespast7years

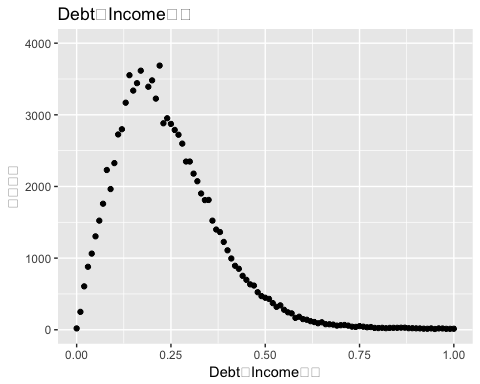
分析发现： 1. 图形为典型的正态分布。 1. Mean = 26.75，Median = 25.00。 1. 贷款的次数还是相当多的，这是一个好的现象。因为这个数据说明了贷款人会多次带宽，能够给公司带来持续的利益。



## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 2.00 17.00 25.00 26.75 35.00 136.00 697

#### v2add - DebtToIncomeRatio

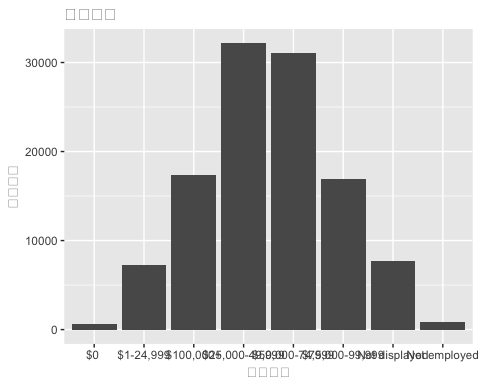
分析发现： 1. 数据中最多比例为10.01，研究了数据集。应该和0收入和低收入的比例有关。 1. 所以此处贷款比例中位数是0.22比平均数更有意义。 1. 但是反思这种情况下本参数对于0收入和低收入的计算没有很大意义。 1. 可以考虑完全将有一定工作收入和没有收入的两类人群拆开，业务也会更有针对性。 1. 为了图形展示，对数量小于5的count和比例大于1的进行了排除。



## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 0.140 0.220 0.276 0.320 10.010 8554

#### IncomeRange

分析发现： 1. 收入数据为明显的正态分布。 1. 这里的not employed和0收入数据的比例不相符，需要额外研究下客户数据发现原因。 1. 感觉分类有些少，区分度不够。



## $0 $1-24,999 $100,000+ $25,000-49,999 $50,000-74,999   
## 621 7274 17337 32192 31050   
## $75,000-99,999 Not displayed Not employed   
## 16916 7741 806

## Part1B - 单变量分析

#### 你的数据集结构是什么？

* 原始数据有81个参数，选择了其中的14个数据，在进行单变量可视化分析的时候有少量排除掉了。结果参见上面一节的说明和图像输出。
* properloan观察。全部113937个数据（第1行是head），在一个80+M的csv中，比较理想。虽然数据量稍大了些，但是经过观察，数据中的变量很多（81个），按照项目要求，选择选择10-15个变量。首先进行数据清理可能会缩减较大尺寸。
* 数据使用的是pls2n，为了处理速度将所选变量的数据存为了csv文件。共有113937个观察，15个变量。并为了处理方便，将空字符变换为了NA。
* 在上到3维分析之后，发现必须处理掉NA数据才能比较直观，又把NA全部处理掉生成了数据集pls2o

str(pls2n)

## 'data.frame': 113937 obs. of 15 variables:  
## $ ListingKey : Factor w/ 113048 levels "00003546482094282EF90E5",..: 7180 7192 6492 6514 6531 6534 6544 6551 6532 6532 ...  
## $ ListingCreationDate : Factor w/ 34801 levels "00:00.0","00:00.1",..: 5431 16498 452 1489 22543 15626 30747 28742 25420 25420 ...  
## $ CreditGrade : Factor w/ 8 levels "A","AA","B","C",..: 4 NA 7 NA NA NA NA NA NA NA ...  
## $ LoanStatus : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4 4 4 4 4 ...  
## $ Term : int 36 36 36 36 36 60 36 36 36 36 ...  
## $ BorrowerAPR : num 0.165 0.12 0.283 0.125 0.246 ...  
## $ ListingCategory : int 0 2 0 16 2 1 1 2 7 7 ...  
## $ Occupation : Factor w/ 67 levels "Accountant/CPA",..: 36 42 36 51 20 42 49 28 23 23 ...  
## $ EmploymentStatusDuration : int 2 44 NA 113 44 82 172 103 269 269 ...  
## $ IsBorrowerHomeowner : logi TRUE FALSE FALSE TRUE TRUE TRUE ...  
## $ TotalCreditLinespast7years: int 12 29 3 29 49 49 20 10 32 32 ...  
## $ DelinquenciesLast7Years : int 4 0 0 14 0 0 0 0 0 0 ...  
## $ 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 ...  
## $ LoanOriginalAmount : int 9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...

* 根据前一节数据的可视化发现：
  + 收入数据、7年内的借款次数为正态分布。
  + 受雇佣的时间为偏斜分布。
  + 带宽类别中，推测默认为1，为Debt Consolidation,从字面解释没有分类意义，可能是默认值。
  + 贷款年化利率呈现的状态为正态分布，中位数为0.20976。最大最小值相差较大。
  + 信用等级评价的分布比较平均，代表中间信用评分的C级最多，其中HR和NC为Posper公司内部分级，通用中无此分类。

#### 你的数据集内感兴趣的主要特性有哪些？

选择这个数据集是对贷款的政策有比较好奇。从数据中可以看到两个方面。首先是ProsperLoan的运行情况。本数据中的贷款的数据很多。从数据的分布可以看出很多属于正态分布，是比较健康的。另外一点，对于贷款人信用的如何评级及贷款的利率等运营决策是否会因贷款人不同而变化，如果有，是怎么变化的比较好奇，这部分将在后两部分进行分析。

#### 你认为数据集内哪些其他特征可以帮助你探索兴趣特点？

初步选择的参数如下，在分析过程中又有一些由于数据和能力问题被放弃了。选择的列表和原因如下： 1. ListingKey 主键 2. （LoanData） 业务视角数据 1. ListingCreationDate 贷款日期 OK 提取年/月 2. CreditGrade 信用等级 OK 虽然缺失较大，但是重要指标，保留，经过观察，和loanstatus有关联，增加指标。 3. LoanStatus 借款状态 4. Term 贷款期（月） 5. BorrowerAPR 年华利率 利率相差较大 6. ListingCategory 行业，数字标识，对应如下： The category of the listing that the borrower selected when posting their listing: 0 - Not Available, 1 - Debt Consolidation, 2 - Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 - Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 - Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 - Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans 。 此处对比Occupation是固定分类，两者有联系，选择这个数据而不是Occupation。 7. Occupation 在职情况 3. （Borrower）借款人视角数据 1. EmploymentStatusDuration 借款人在职时间 数据较全（考虑到房产的情况，增加1个指标） 2. IsBorrowerHomeowner 借款人是否有房产 3. TotalCreditLinespast7Years 7年内借款次数 数据较全 4. DelinquenciesLast7Years 7年内欠款次数 数据较全 5. DebtToIncomeRetio 债务与收入比例 数据较全 6. IncomeRange 收入范围 数据较全 7. （ProsperPaymentsLessThanOneMonthLate 一月内还清带宽标识）区分性较小 8. （PercentFunded 贷款比例）基本都是1，区分性小 9. （Recommandations 推荐人）区分性小

#### 根据数据集内已有变量，你是否创建了任何新变量？

没有创建新的变量。

#### 在已经探究的特性中，是否存在任何异常分布？你是否对数据进行一些操作，如清洁、调整或改变数据的形式？如果是，你为什么会这样做？

在统计时发现is.na无法正常过滤，经过检查大部分的空值是空字符而不是NA。调整了读入csv的命令，经过测试，问题解决：

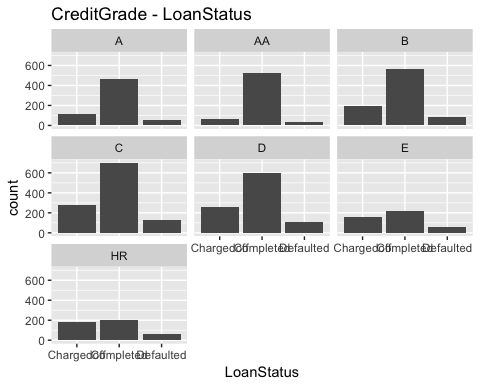
pls2n <- read.csv(‘prosperLoanData\_s.csv’, na.strings=c(“”, “NA”))

另外在一些异常值的问题上，按照百分比进行了限制了Data = pls2n的显示效果，没有选择写入到文件中的方法。

## Part2A - 双变量绘图选择

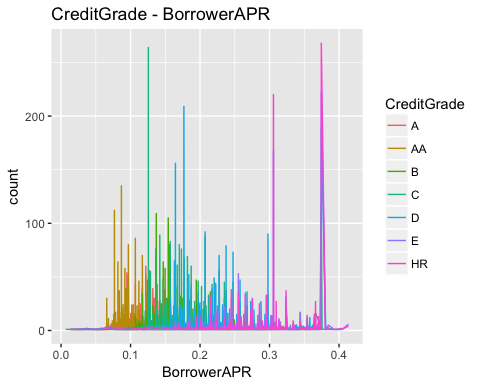
#### CreditGrade - LoanStatus

可以看出，A和AA的完成比例最高。而最后两个信用档则比例非常低。以上数据说明了评级还是比较准确的（也可能LoanStatus本身计算参数就含LoanStatus）。 另：为了排斥facet中NA的影响，此处使用的是通过omit去掉所有NA值的pls2o数据，观察值为20225个。



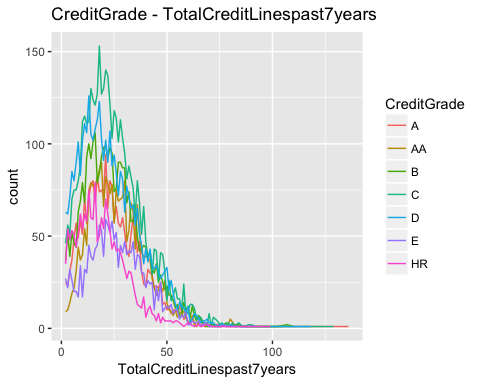
#### CreditGrade - BorrowerAPR

根据图形所示，非常明显信用等级越高，年化利率就越低，分布十分理想。另外HR的分类特别的高，而且跨度超长，可能和定义有关，愿数据说明没有此部分细节。



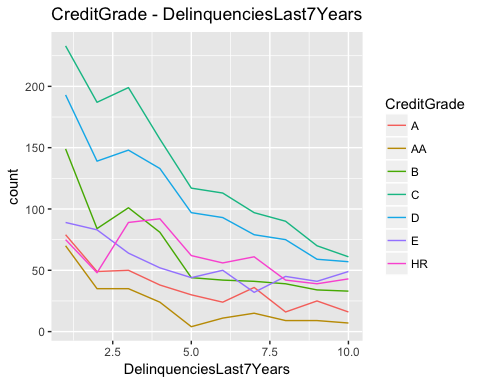
#### CreditGrade - TotalCreditLinespast7years

根据图形显示，低等级的评分的贷款数量明显大于高等级的数量。说明低等级信用贷款在数量上是十分重要的。并且贷款数量与信用等级相关性基本没有。



#### CreditGrade - DelinquenciesLast7Years

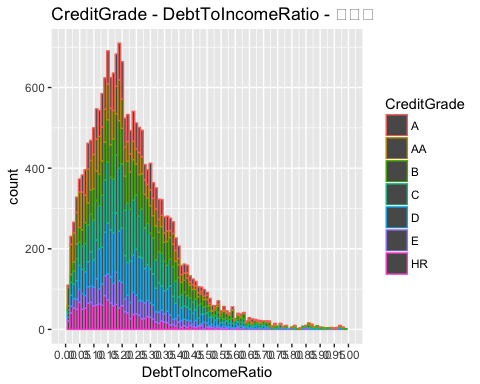
根据图形显示（为了便于观察，对x做了限制），信用等级和超期还款有强相关性。



#### CreditGrade - DebtToIncomeRatio - his

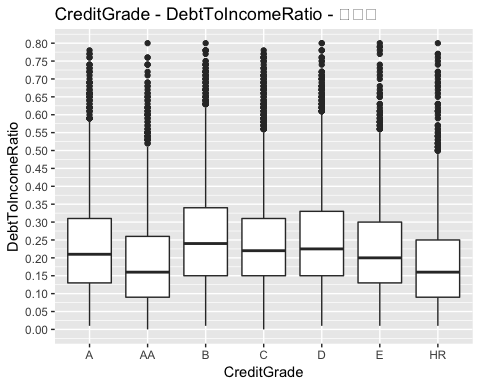
根据图形显示（为了便于观察，对x做了限制），借款人的偿还能力对信用等级并没有影响。贷款利率的众数是0.19。

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1300 0.2100 0.3425 0.3200 10.0100



#### CreditGrade - DebtToIncomeRatio -box

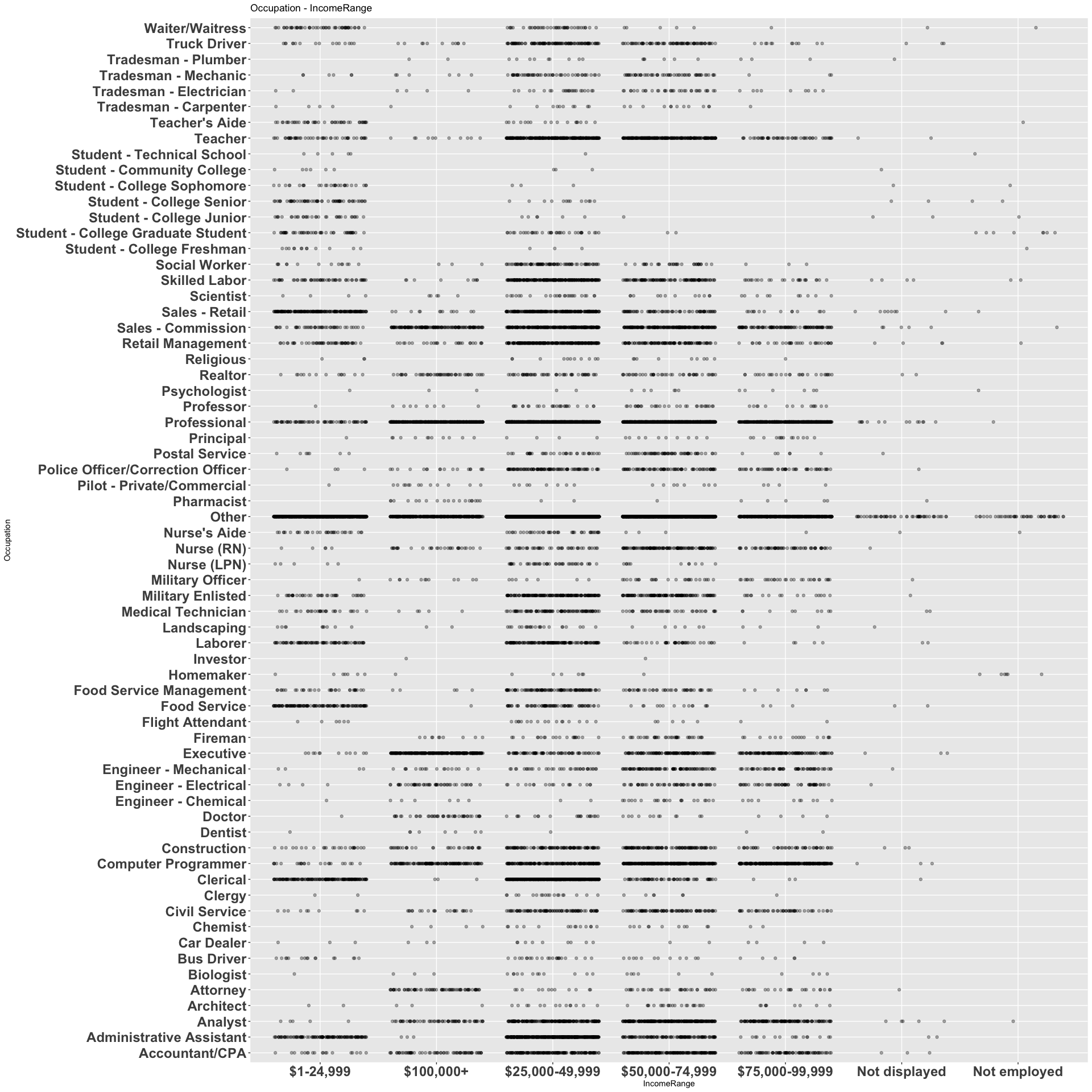
对比上一个图，箱线图更容易显示出数据的对比。



#### Occupation - IncomeRange

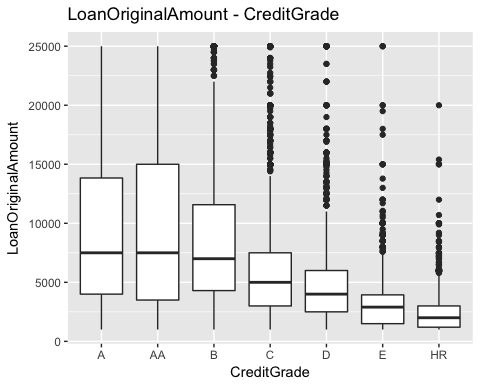
根据图形显示观察Occupation和IncomeRange的关系。发现数据特别分散，实际意义不大。而且还有个Other分类聚集了大部分数据。

##   
## $0 $1-24,999 $100,000+ $25,000-49,999 $50,000-74,999   
## 0 2511 2037 7823 5284   
## $75,000-99,999 Not displayed Not employed   
## 2354 139 77



#### LoanOriginalAmount - CreditGrade

根据图形显示贷款评级越高平均贷款额就越高，两个参数相关。



## Part2B - 双变量分析

#### 探讨你在这部分探究中观察到的一些关系。这些感兴趣的特性与数据集内其他特性有什么区别？

本部分观察的重点是CreditGrade到底是和那些指标有关系。之后又好奇Proper收集的IncomeRange数据和Occupation之间有没有展示一些信息，这一部分由于数据比较分散，且有Other的默认选项，没有很大价值，推荐将一些分类合并，比如Traindsman，Teacher可以代表相关的两大类别。并降低Other的选项比例。

#### 你是否观察到主要特性与其他特性之间的有趣关系？

CreditGrade的评级还是非常有用的指标，对于LoanStatus的区分对应比较好。比较好玩的是，这个指标和贷款次数没有关系，只和贷款违约有关系。

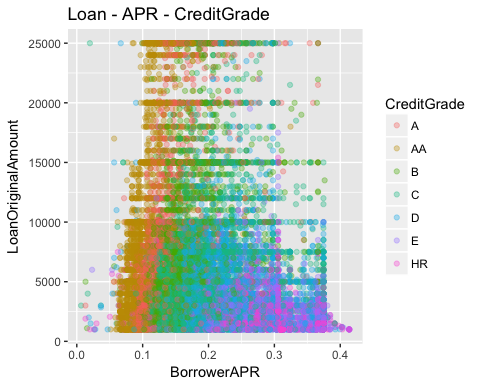
#### 你发现最强的关系是什么？

通过图形判断。CreditGrade和以下指标具备相关性：DelinquenciesLast7Years，BorrowerAPR。

## Part3A - 多变量绘图选择

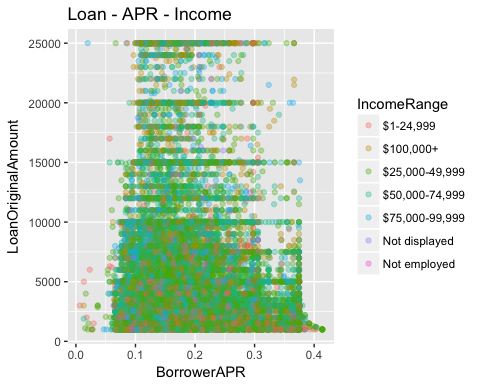
#### Loan - APR - CreditGrade

根据图形显示PorsperLoan对高信用的用户青睐有加，更低的利率，更多的金额。



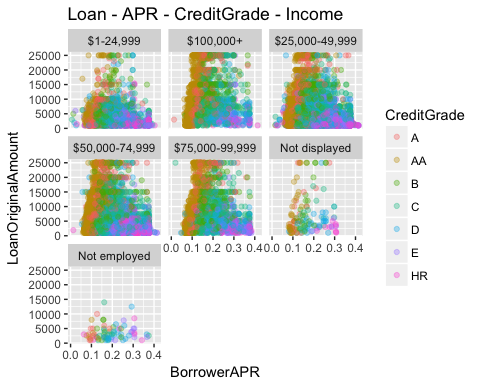
#### Loan - APR - Income

根据图形显示PorsperLoan对IncomeRange倒是没什么特别关注。



#### Loan - APR - CreditGrade - Income

用Facet增加第4纬度后，分离Income后还是可以看出虽然和额度关系不大，但和是否能够贷款审批关系比较重大。



## Part3B - 多变量分析

#### 探讨你在这部分探究中观察到的一些关系。通过观察感兴趣的特性，是否存在相互促进的特性？

主要观测的是贷款额-利率的关系，因为这个和公司经营十分密切。又增加了信用纬度和收入纬度，这两方面涉及公司的风险控制。

#### 这些特性之间是否存在有趣或惊人的联系呢？

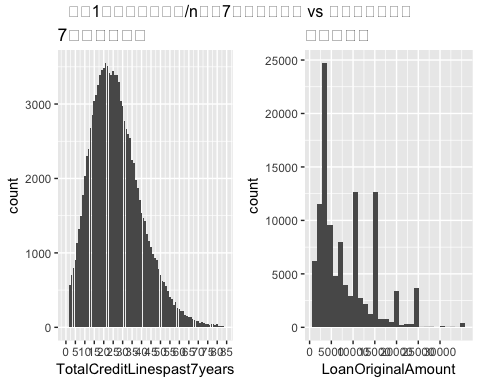
发现贷款额和信用正相关，利率和信用负相关。有趣的是收入制作为准入门槛，和额度及利率关系不大。

#### 选项：你是否创建过数据集的任何模型？讨论你模型的优缺点。

贷款的数据比钻石的价格数据更复杂，变量达到了81个，经过尝试，找到了一些相关的关系。其中最重要的指标是Credit，和许多其他指标都有关系。但是比较遗憾的是尝试了比较多的转换，未找到类似钻石线形关系的对应关系。

## Part4 - 定稿图与总结

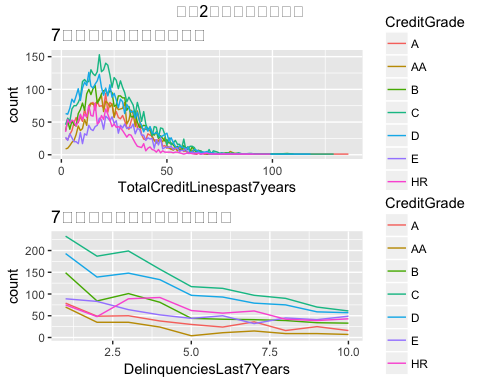
#### 绘图一



#### 描述一

绘图1为单一纬度的两个探索的聚合，左边是Prosper公司7年内发生借款的数量，符合正态分布。右侧是贷款额的统计（不是正态分布，但数据比较健康）。从这两个指标可以看出Posper公司的经营情况比较正常。

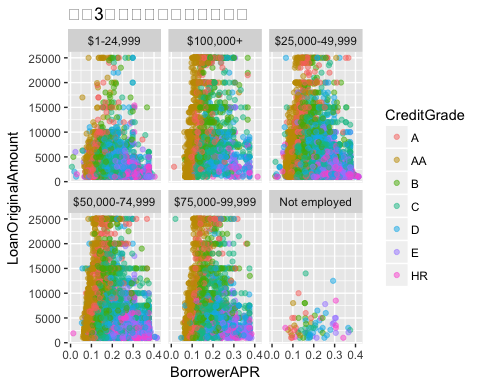
#### 绘图二



#### 描述二

绘图2为二纬度两个探索的聚合，上面是贷款次数和信用的关系，下面是违约还款和信用的关系。可以看出信用和违约还款次数有较强负相关，但和贷款次数关系不大。

#### 绘图三



#### 描述三

用Facet增加第4纬度后，分离Income后还是可以看出虽然和额度关系不大，但和是否能够贷款审批关系比较重大。从这个分析可以看出收入水平是是否可以贷款的准入条件，但是一旦满足门槛，关系就不那么大了。

## Part5 - 反思

项目大概做了5天，由于选择的是和课程钻石价格预测相差较大的数据，做的比较苦手。最终选择的是properloan。酒类的不是很感兴趣，而且数据量比较小。总统选举的不感兴趣。贷款的不错，其中paipaidi的都是excel数据，仔细看了一遍说明和数据文件。多个数据文件比较繁琐，而且格式不友好（有个表的中文都是乱码）还有个表的把期数作为primarykey无法理解。需要额外的工作比较多。

在处理的过程中，感觉到数据分析不只是做清理、数据处理和可视化。有很大部分是对各个参数的理解，参数之间关系的估计和验证。和这些验证之后要达到目标的分析。这些可能才是探索性数据分析的难点。

在处理过程中，有以下问题，情评审老师解答： 1. 在最后一副图中，怎样把facet中的‘Not displayed’分类去掉，即有选择的进行facet展示？