

R programlama diliyle veri seti incelemesi: Prosper.com kredi verileri

Bu çalışmam, R programlama diliyle veri seti analizine örnek bir uygulamadır, aynı zamanda Udacity - "Data Analysis with R" dersi için de yaptığım final projesidir. Veri seti csv formatında olup 82 değişken ve 113937 gözlemden oluşmaktadır. "**ProsperLoanData**" adındaki bu veri setine ilk göz atışım sonrası değişkenlerin neyi ifade ettiğine dair ön bir araştırma yaptım, öyle ki veriler Türkiye'de benzeri olmayan bir iş modeline ait. *Prosper.com* adıyla ABD'de faaliyet gösteren bu internet sitesi, basit bir tanımla kredi ihtiyacı olanların ve kredi veren bireysel yatırımcıların buluştuğu bir platform. Veri setindeki her bir gözlem de, gerçekleşen kredi işlemine dair bilgileri barındırmaktadır. Bu paylaşım, R Studio platformunda bir veri setinin analiz sürecinde *ggplot2*, *dplyr*, *tidyverse* başta olmak üzere kullanılan kütüphane ve kodları içermektedir. Bu paylaşımmdaki amacım, R diline ilgi duyanlar veya öğrenmeye yeni başlayanlar için merak ve kimi sorularını giderebilecekleri bir alıntıma örneği sunmaktır. Bir alıntıma uygulaması olması nedeniyle herhangi bir sorunun çözümüne yanıt aramaktan ziyade, bu çalışmada R dilinde öğrenilenlerin pratikte uygulanması amacıyla güdüldüğünden çözümlemeler serbest bir yaklaşımla yapılmıştır.

```
#Verinin yüklenmesi ve
#ilgili kütüphanelerin yüklenmesi
setwd("C:/Users/User/Downloads/")
ld <- read.csv("prosperLoanData.csv")
library(dplyr)
library(tidyverse)
library(ggplot2)
```

İlk olarak değişkenler içinde analiz için gereksiz görülenler tespit edilerek veri setinden çıkarılmıştır.

```
#Gereksiz görülen/istenmeyen değişkenlerin veri setinden çıkarılması
ld$ListingKey <- NULL
ld$ListingNumber <- NULL
ld$LoanKey <- NULL
```

Harf ile ifade edilen kredi notları("*CreditGrade*") ve Prosper derecelendirme ("*ProsperRating..Alpha.*") değişkenlerinin grafiklerde rahatlıkla anlaşılması için harfler en yüksekten en düşük değeri temsile göre sıralanmıştır. Aynı amaçla gelir grubu ("*IncomeRange*") değişkeni de düzenlenmiştir.

```
#Değişken değerlerinin belirtilen bir sıraya göre dizilmesi
ld$CreditGrade <- ordered(ld$CreditGrade, levels = c("AA", "A", "B", "C", "D", "E", "HR", "Nc", "NA"))
ld$ProsperRating..Alpha. <- ordered(ld$ProsperRating..Alpha., levels =
                                         c("AA", "A", "B", "C", "D", "E", "HR", ""))
ld$IncomeRange <- ordered(ld$IncomeRange, levels = c("Not displayed", "Not employed", "$0", "$1-24,999"))
```

separate() fonksiyonu kullanılarak "*LoanOriginateDate*" değişkeni içinde yer alan değerler yıl("*L.O.Year*"), ay("*L.O.Month*") ve gün("*L.O.Day*") olarak ayrılarak yeni değişkenler oluşturulmuştur.(Saat, dakika ve saniye kısımları tamamen çıkarılmıştır.)

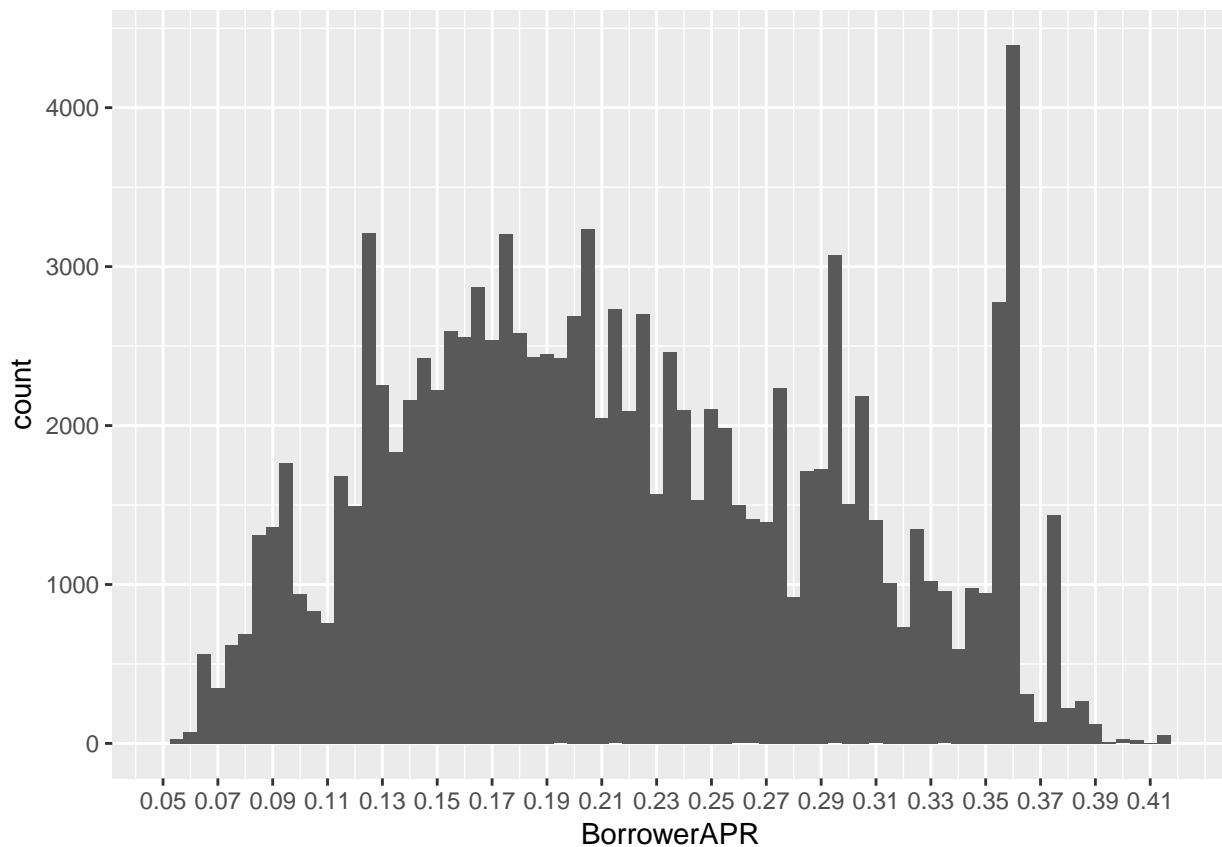
```
#Değişken değerleri karakterlerine göre ayırlarak yeni değişkenler oluşturulması
ld <- separate(ld, LoanOriginationDate, c("L.O.Year", "L.O.Month", "L.O.Day" ), sep = "-")
##"L.O.Day" değişkeni içinde yer alan 'gün, saat, dakika, saniye' değerleri içinde sadece gün değerinin
ld$L.O.Day = unlist(strsplit(ld$L.O.Day, split=" ", fixed=TRUE))[1]
```

Tek değişkenli görseller

Merak ettiğim ilk husus bankalara göre daha avantajlı faiz oranı sunma iddiasındaki Prosper'in faiz oranlarını görselleştirmek oldu. Öncelikle `summary()` fonksiyonu ile edindiğim özet bilgiye göre veri setinde yer alan kredi faiz oranlarının en düşüğü 0.00653 iken en yüksek 0.51229 oranı olduğu görülmüştür. Medyan fazi oranı değeri 0.20976, ortalama faiz oranı ise 0.21883'dür. Histogram grafik ile görselleştirildiğinde dağılm 0.40 oranından sonra çok düşük olduğu için ggplot kodlarından gözlemleneceği üzere 0.05-0.42 değerleri aralığında sınırlandırılmıştır.

```
#Kredi faiz oranlarının dağılımı
summary(ld$BorrowerAPR)
```

```
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## 0.00653 0.15629 0.20976 0.21883 0.28381 0.51229      25
ggplot(data = ld, aes(x = BorrowerAPR)) +
  geom_histogram(binwidth = 0.005) +
  scale_x_continuous(limits = c(0.05, 0.42), breaks = seq(0.05, 0.42, 0.02))
```



Tahmin etmek zor olmasa da görselden görüldüğü üzere; Prosper'da kredi alanlar içinde iş durumlarına(“Employment Status”) göre çalışanlar (“Employed” ve “Full-time”) ağırlıklıdır. Çalışmayanlar(“Not employed”), serbest çalışanlar(“Self-employed”), kısmi zamanlı çalışanlar(“Part-time”), emekliler(“Retired”) ve çalışma durumu hakkında bilgisi bulunmayan kredi borçları (“Not available”, “Other” ve boş) görece azınlıktadır.

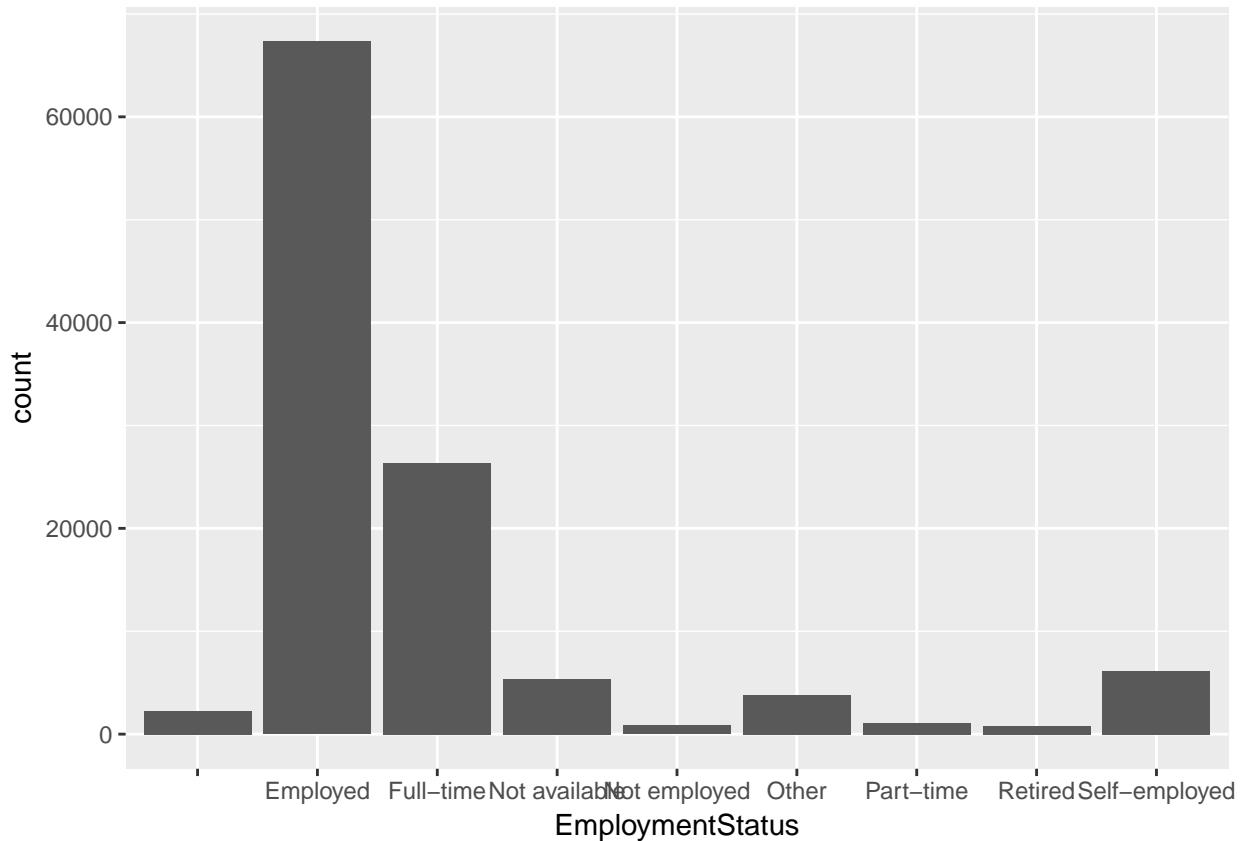
```
#Kredi borçlarının iş durumları
summary(ld$EmploymentStatus)
```

	Employed	Full-time	Not available	Not employed
##	2255	67322	26355	5347
##				835

```

##          Other      Part-time       Retired Self-employed
##          3806        1088         795           6134
ggplot(l1d, aes(EmploymentStatus)) + geom_bar()

```



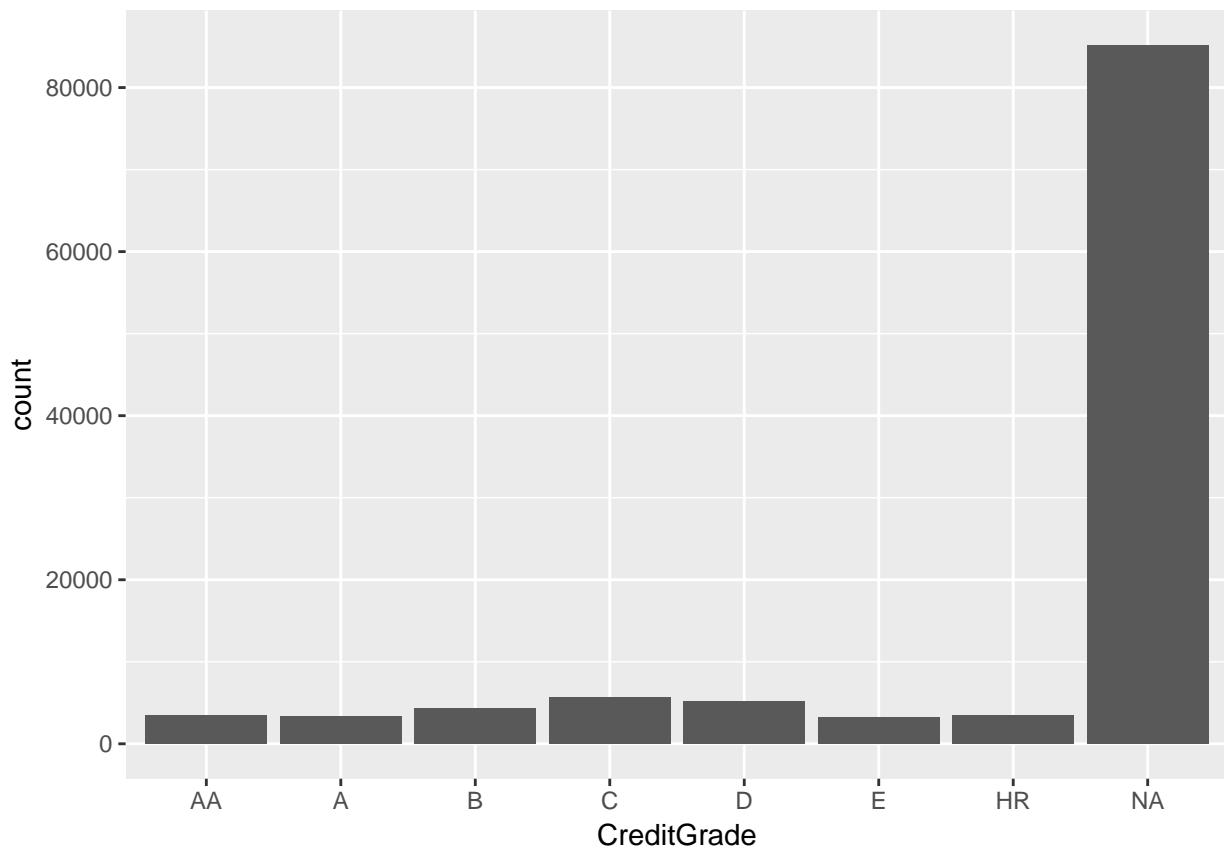
Kredi notlarına göre borçluların dağılımı incelendiğinde herhangi bir nota sahip olmayan borçluların çok daha fazla olduğu görülmüştür. Çoğunluğu oluşturan bu grup çıkarıldığında kredi notlarına göre borçlular normal dağılım göstermektedir.

#Kredi notlarına göre borçluların dağılımı
`summary(l1d$CreditGrade)`

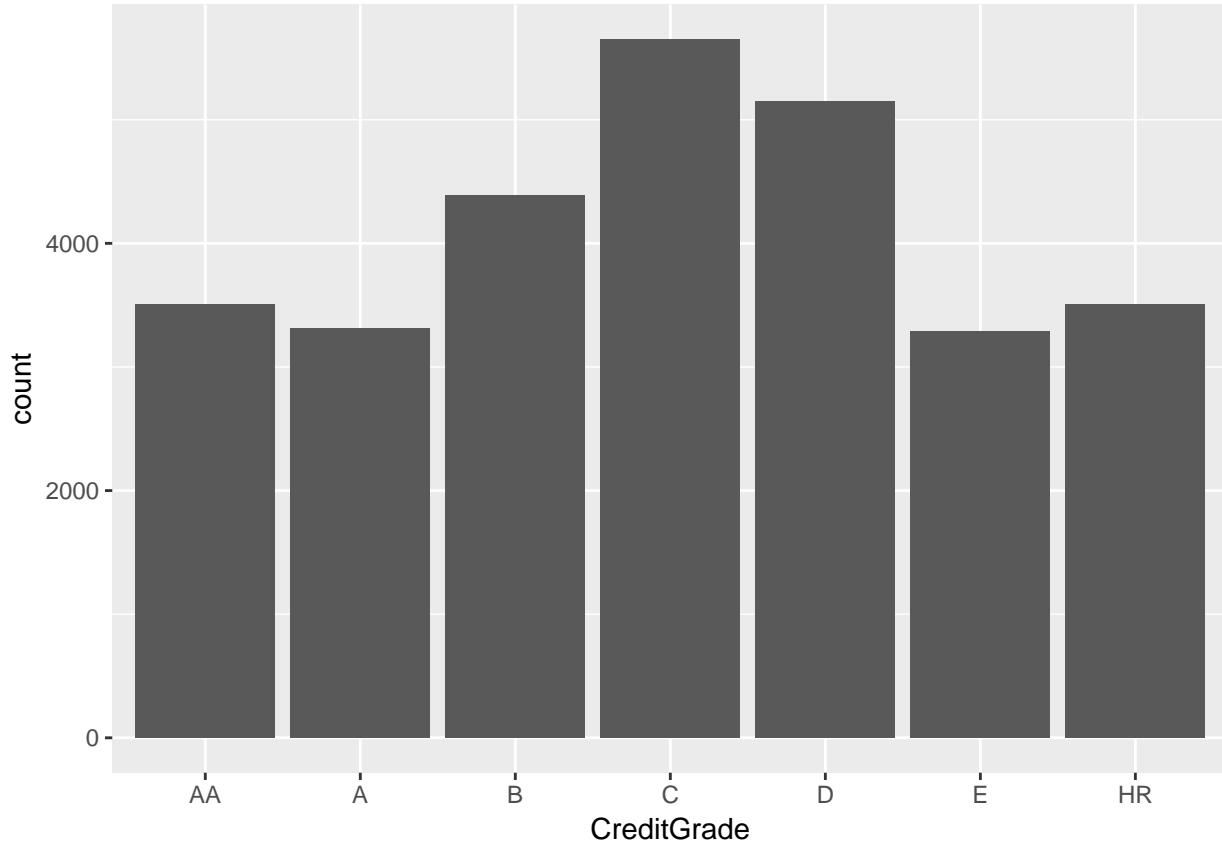
```

##    AA      A      B      C      D      E      HR     Nc     NA   NA's
##  3509  3315  4389  5649  5153  3289  3508     0     0  85125
ggplot(data = l1d, aes(CreditGrade)) + geom_bar()

```



```
#Görselden kredi notuna sahip olmayan gözlemlerin çıkarılması  
ggplot(data = subset(ld, ld$CreditGrade != ""), aes(CreditGrade)) + geom_bar()
```

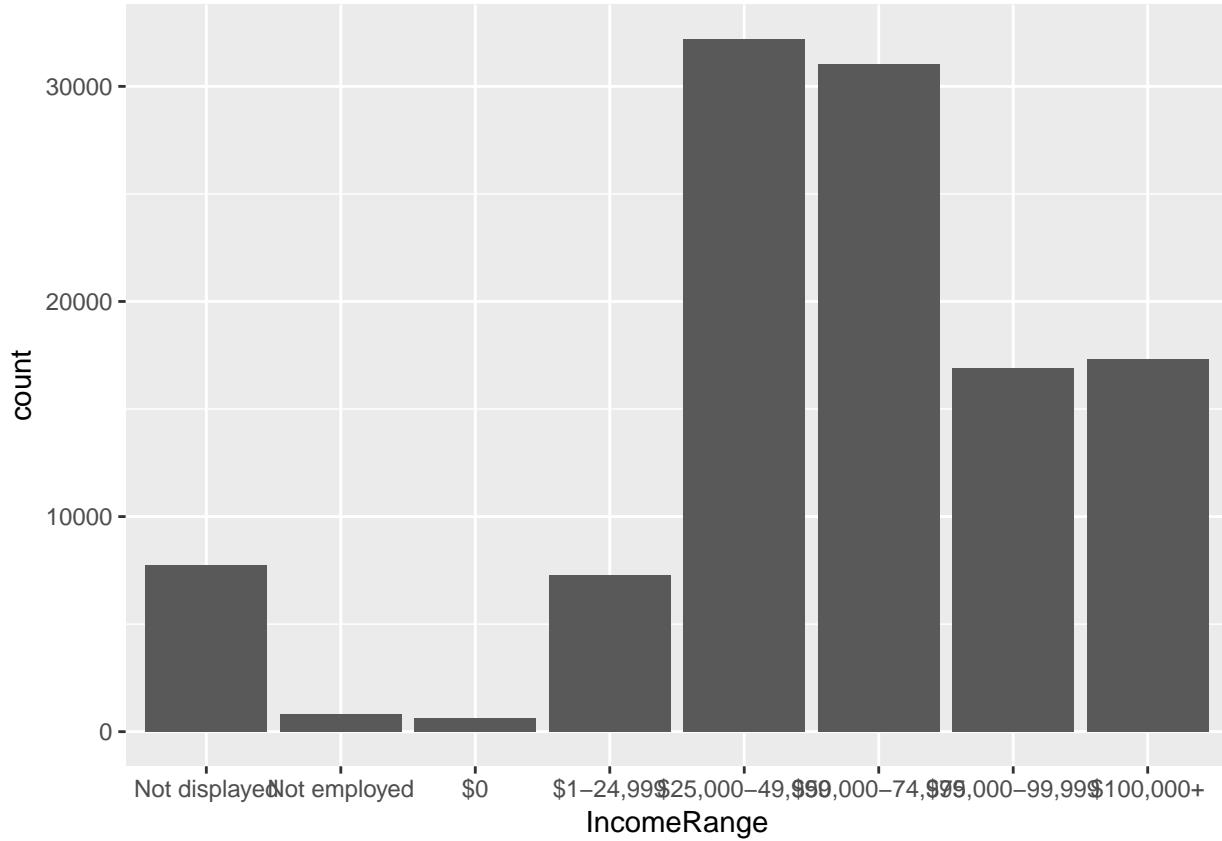


Veri setinde yer alan, gelir gruplarına göre Prosper üzerinden kredi alanların yarısını 25-50 bin ile 50-75 bin dolar yıllık gelire sahip borçlular oluşturmaktadır. Bu iki gruptaki borçluların sayısı, 75-100 bin ile 100 bin üzeri gelir grubuna dahil borçluların sayısının yaklaşık ikişer katıdır. Herhangi bir gelir grubuna dahil görünmeyen/gelir beyan etmeyen (“Not Displayed”, “Not employed” ve “\$0”) ve yıllık gelirleri 1-25 bin dolar aralığında olan borçluların sayısı toplamı da üst gelir grubuna (75-100 ve 100+) dahil borçlu sayısının yarısı civarındadır.

```
#Kredi borçlularının gelir gruplarına göre dağılımları
summary(ld$IncomeRange)
```

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

```
ggplot(data = ld, aes(IncomeRange)) + geom_bar()
```



Prosper.com üzerinden kredi kullananların, kredi çekme nedenleri de merak ettiğim bir diğer konuydu. Çubuk grafik üzerinde beyan edilen nedenlerin dağılımını göstermeden önce veri setinde sayısal olarak yer alan değerleri grafikte metin biçiminde göstermek için 1-20 aralığında değişen her biri ayrı nedeni temsil eden sayısal değerler değişken tanımlamalarının yer aldığı bilgilendirme notuna göre metin biçiminde düzenlenmiştir. Görüldüğü üzere 133937 kredi borçlusuna ait verilerin yarısından fazlası için kredi çekme nedeni beyanında bulunulmamıştır (“Not Available” + NA). Mevcut beyanlara göre, otomobil(“Auto”), borç transferi(“Debt Consolidation”) ve ev harcamaları(“Home Improvement”) amacıyla çekilen kredilerin sayısı öne çıkmaktadır. Grafikte “Wedding Loans” değerine ait bir çubuk görmeyince `table()` fonksiyonunu kullanarak bu değere ait bir veri olmadığını gördüm.

```
#Kredi çekme nedenlerinin sayısal dağılımı
count(ld, ListingCategory..numeric.)
```

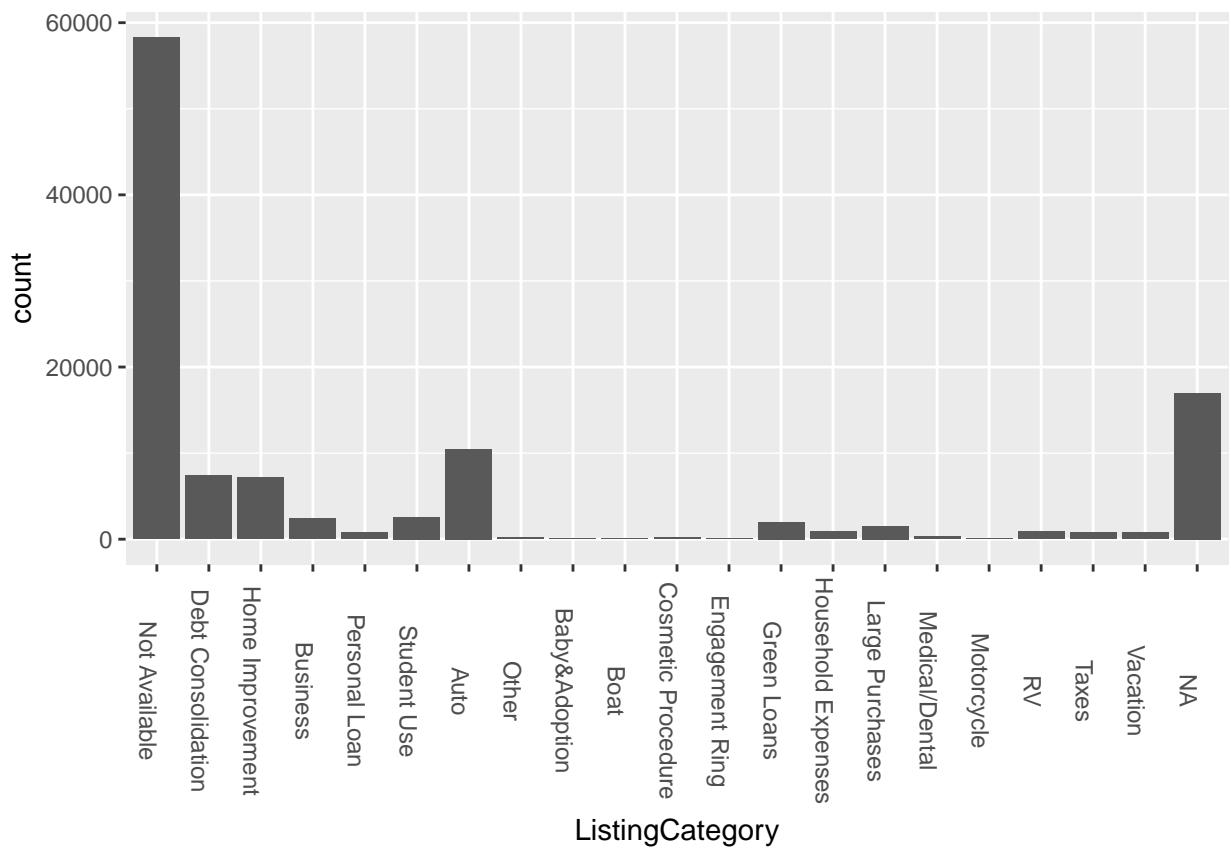
```
## # A tibble: 21 x 2
##   ListingCategory..numeric.     n
##   <int> <int>
## 1 0      16965
## 2 1      58308
## 3 2      7433
## 4 3      7189
## 5 4      2395
## 6 5      756
## 7 6      2572
## 8 7      10494
## 9 8      199
## 10 9      85
## # ... with 11 more rows
```

```

#Kredi çekme nedenlerini temsil eden sayısal değerlerin metine dönüşümü
x <- c("Not Available", "Debt Consolidation", "Home Improvement",
      "Business", "Personal Loan", "Student Use", "Auto", "Other",
      "Baby&Adoption", "Boat", "Cosmetic Procedure", "Engagement Ring",
      "Green Loans", "Household Expenses", "Large Purchases",
      "Medical/Dental", "Motorcycle", "RV", "Taxes", "Vacation",
      "Wedding Loans")
ld$ListingCategory <- factor(ld$ListingCategory..numeric.,
                               levels = seq(0:20), labels = x)

#Kredi çekme nedenlerinin dağılımı
ggplot(data=ld, aes(x=ListingCategory)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = -90))

```



```
table(ld$ListingCategory)
```

```

##          Not Available Debt Consolidation   Home Improvement
##                  58308                      7433                     7189
##          Business        Personal Loan     Student Use
##                  2395                      756                     2572
##          Auto           Other         Baby&Adoption
##                  10494                     199                     85
##          Boat        Cosmetic Procedure Engagement Ring
##                  91                       217                     59
##          Green Loans Household Expenses Large Purchases

```

```

##          1996          876          1522
## Medical/Dental      Motorcycle      RV
##          304           52           885
## Taxes            Vacation       Wedding Loans
##          768           771           0

```

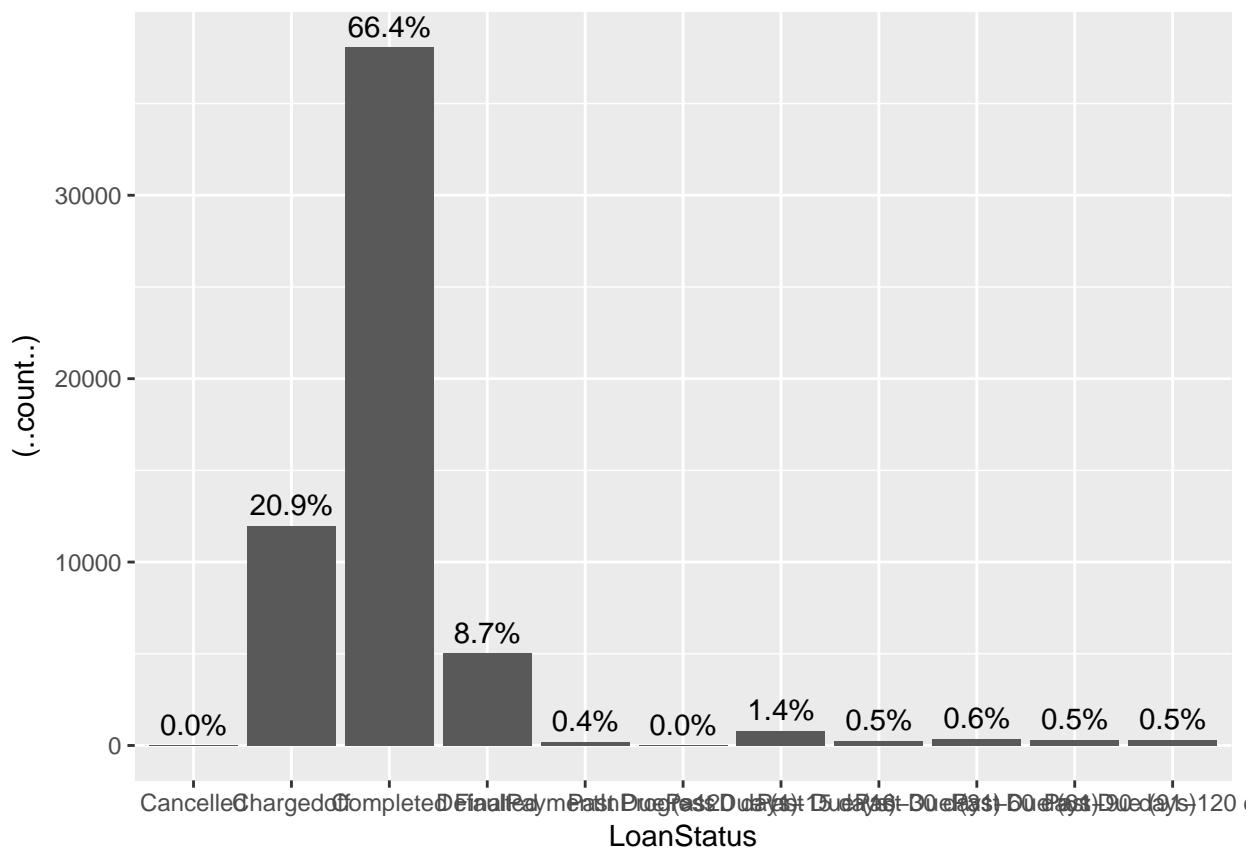
Prosper'a yatırım yapacak bir yatırımcının belki de ilk merak edeceğii sorunun yanıtını aradım; "Bugüne dek gerçekleşmiş kredi işlemlerinde hesap durumlarının dağılımı nedir?". Bunun için, akibetinin ne olacağı bilinmediği için verisetinde ödemesi halihazırda devam eden - "Current" durumundaki- hesaplar çıkarılarak sonuçlar çubuk grafik üzerinde gösterilmiştir. Elde edilen çıktıya göre ödemesi tamamlanan hesapların oranı 2/3 oranındadır. Diğer bir deyişle gecikmiş ödeme veya ödenmeyen borçlardan ötürü yatırımcıların içte bir ihtimalle baş ağrısı çekme olasılığı vardır.

```

library(scales)

ggplot(data = subset(l1d, LoanStatus != 'Current'), aes(x = LoanStatus)) +
  geom_bar(aes(y = ..count..)) +
  geom_text(aes(y = ..count..), vjust = -0.5, label = ifelse(..count.. == 0, "", scales::percent(..count..)))

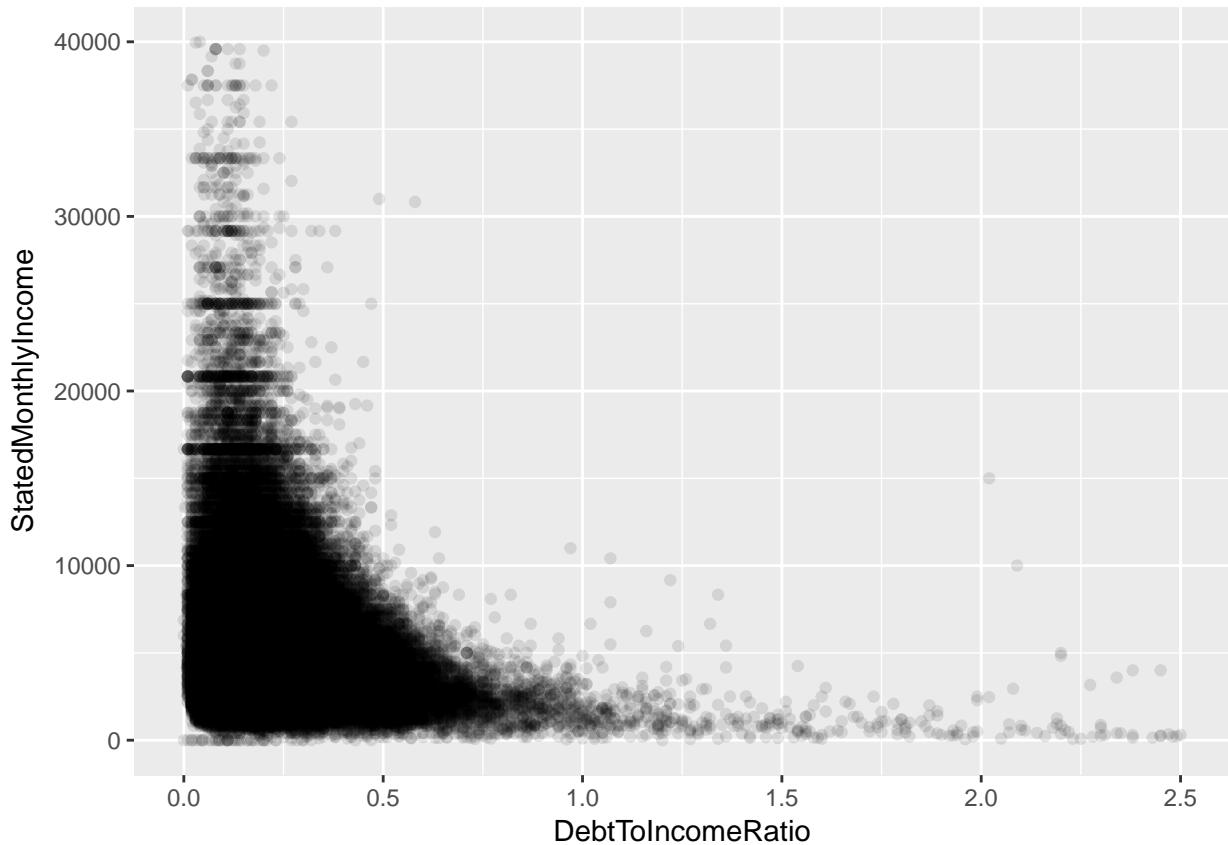
```



İki değişkenli görseller

Borç/gelir oranının aylık gelire göre dağılımı aşağıda gösterilmiştir. Yoğunluk aylık 0-10 bin dolar gelir aralığıyla 0.5 borç/gelir oranı arasındayken, yüksek borç/gelir oranına sahip borçluların düşük aylık gelire sahip oldukları görülmektedir.

```
#Borç/gelir oranının aylık gelir ile karşılaştırılması
ggplot(data = ld, aes(x=DebtToIncomeRatio, y=StatedMonthlyIncome)) +
  geom_jitter(alpha = 1/10, position = position_jitter(h = 0)) + xlim(0, 2.5) + ylim(0, 40000)
```



Prosper.com verilerine göre mesleklerde göre aylık ortalama gelirleri karşılaştırmak için bu iki değişkeni görselleştirdiğimde veri setindeki meslek çeşitliliğinin fazla olması nedeniyle x ekseninde tüm meslek adlarının birbirine girerek okunması mümkün olmayan bir grafik ortaya çıktı. Meslekleri incelediğimde bazı mesleklerin bireleştirilerek tek bir değişken olarak ifade edilebileceğini fark ettim. Uygulamada sadece öğrenci grupları bireleştirilerek "Student" adlı yeni bir değişken yaratılmıştır.

```
#Meslek ('Occupation') değişkeni değerlerinden farklı gruplardan öğrencilerin
# "Student" adlı tek bir değere eşitlenmesi
job_groups <- group_by(ld, Occupation)
ld.by_jobs <- summarise(job_groups,
                         monthly_income_mean = mean(StatedMonthlyIncome), monthly_income_median = median(StatedMonthlyIncome))
ld.by_jobs <- t(ld.by_jobs)
colnames(ld.by_jobs) <- ld.by_jobs[1, ]
ld.by_jobs <- ld.by_jobs[-1, ]
colnames(ld.by_jobs)[1] <- ""
ld.by_jobs = as.data.frame(ld.by_jobs)
ld.by_jobs[,1:68] <- lapply(ld.by_jobs[,1:68], function(x) as.numeric(as.character(x)))
ld.by_jobs$Student <- rowMeans(ld.by_jobs[,54:60], na.rm = TRUE)
ld.by_jobs <- ld.by_jobs[,c(1:53, 61:69)]
ld.by_jobs <- ld.by_jobs[,c(1:53, 62, 54:61)]
ld.by_jobs <- t(ld.by_jobs)
ld.by_jobs = as.data.frame(ld.by_jobs)
```

```

library(data.table)
ld.by_jobs <- setDT(ld.by_jobs, keep.rownames = TRUE) []
colnames(ld.by_jobs)[1] <- "Occupation"

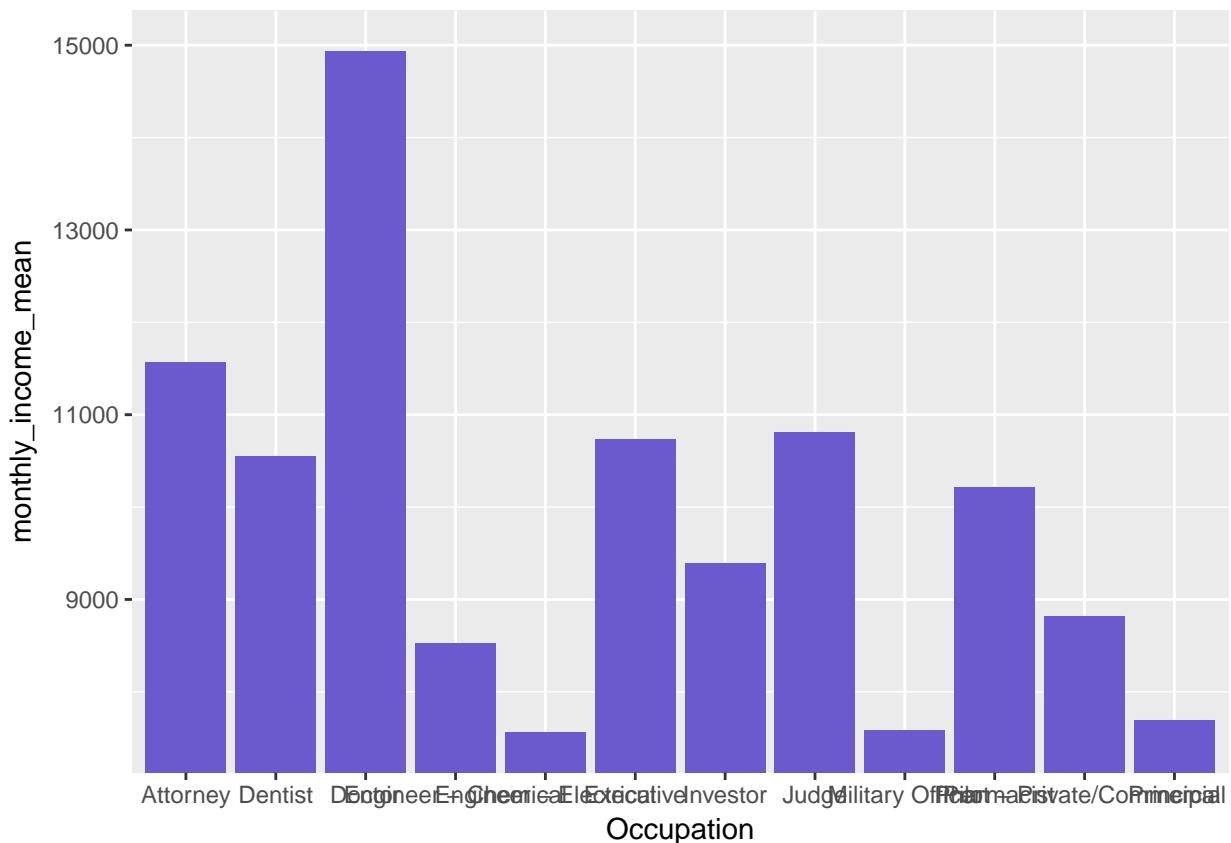
```

Bu işlem ile mesleklerin aylık gelir ortalaması ve aylık gelir medyan değerleri ve frekanslarını barındıran “ld.by_jobs” adlı data.frame içerisinde “Occupation” değişkeni değerlerinden “Student-College Freshman”, “Student-College Graduate Student”, “Student-College Junior”, “Student-College Senior”, “Student-College Sophomore”, “Student-Community College” ve “Student - Technical School” değerleri oluşturulan “Student” değeri altında toplanmıştır. Mevcut 62 meslek içinde aylık ortalama geliri 7500 dolar üzerinde olan meslekleri görselleştirmeyi tercih ettim. Buna göre bu sınırı geçen 12 meslek grubu görülmektedir.

```

#Aylık ortalama geliri 7500 üzerinde olan meslekler
ggplot(data = subset(ld.by_jobs, ld.by_jobs$monthly_income_mean>7500), aes(x = Occupation, y = monthly_
  geom_bar(stat = "identity", fill="slateblue") + coord_cartesian(ylim=c(7500,15000))

```



En iyi kredi notuna sahip mesleklerin dağılımında en az 100 “AA” notuna sahip meslekler, %>% operatörüyle select(), filter(), group_by(), arrange() ve summarise() fonksiyonları birlikte kullanılarak gösterilmiştir.

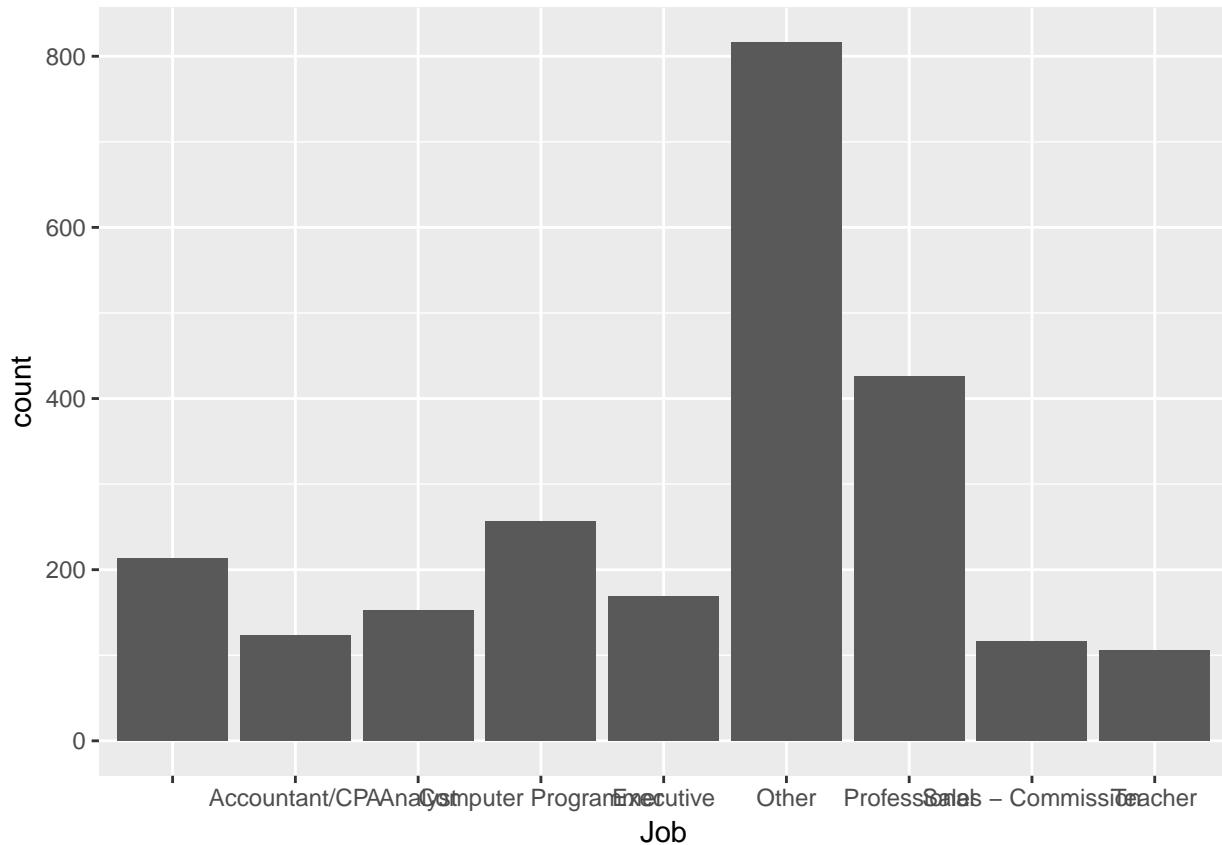
```

#Sayısı en az 100 olan, "AA" kredi derecesine sahip borçluların meslekleri
Jobs_w_AA_Grade <- ld %>%
  select(Job = Occupation, Grade = CreditGrade) %>%
  filter(Grade == "AA") %>%
  group_by(Job) %>%
  arrange(Job) %>%
  summarise(count = n()) %>%
  filter(count >= 100)

```

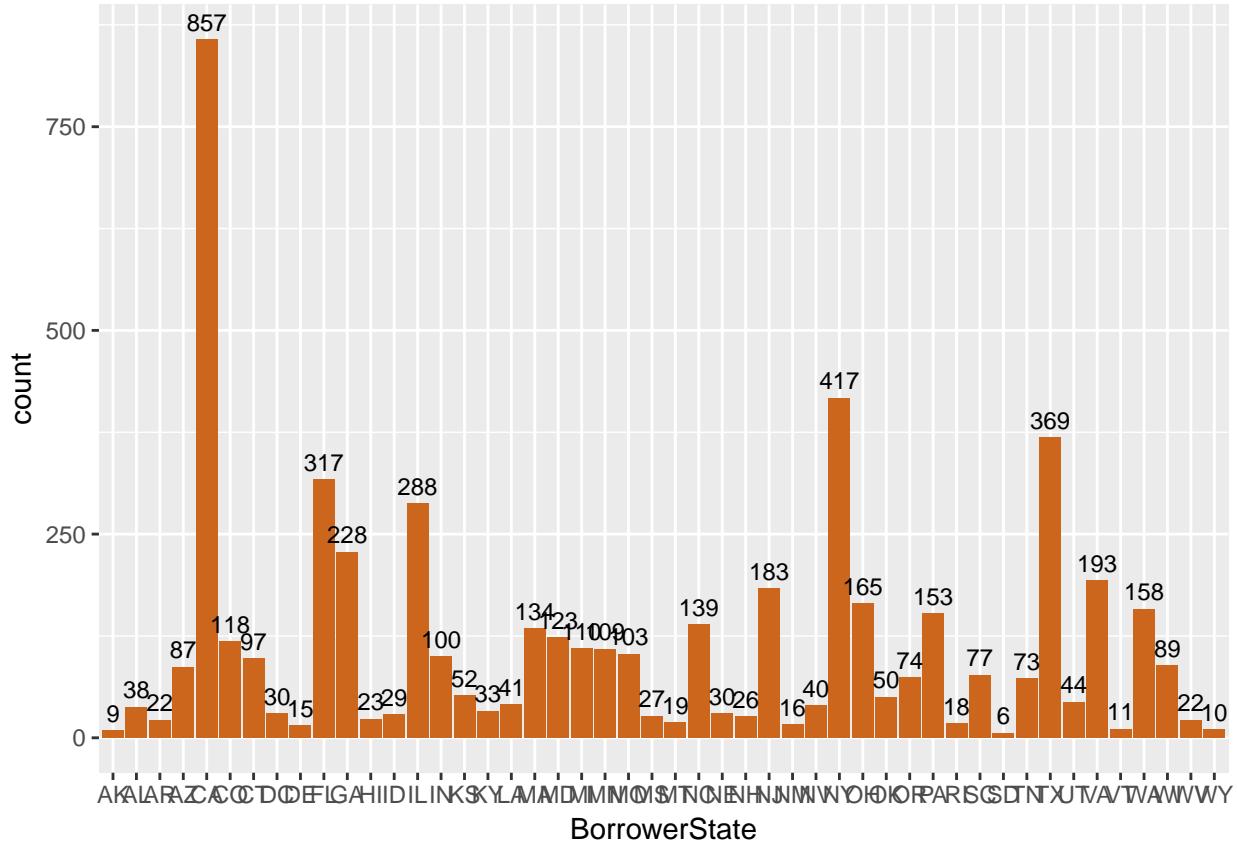
```
#'Matrix' formatının 'data.frame'e dönüştürülmesi
Jobs_w_AA_Grade <- as.data.frame(Jobs_w_AA_Grade)
```

```
ggplot(data = Jobs_w_AA_Grade, aes(x = Job, y = count)) +
  geom_bar(stat = 'identity')
```



Aşağıdaki grafik tek değişkene ait bir görsel olsa da, yine “AA” kredi notuna sahip borçlulara ait bir veri ve `subset()` fonksiyonunun kullanımından ötürü bu bölümde yer almıştır. Grafikte “AA” Prosper derecelendirme notuna sahip borçluların eyaletlere göre dağılımları görülmektedir. Borçlular arasında AA notuna sahip olanların bulunduğu eyaletler içinde Prosper.com'un adresi olan Kaliforniya(CA) eyaleti açık ara öndedir.

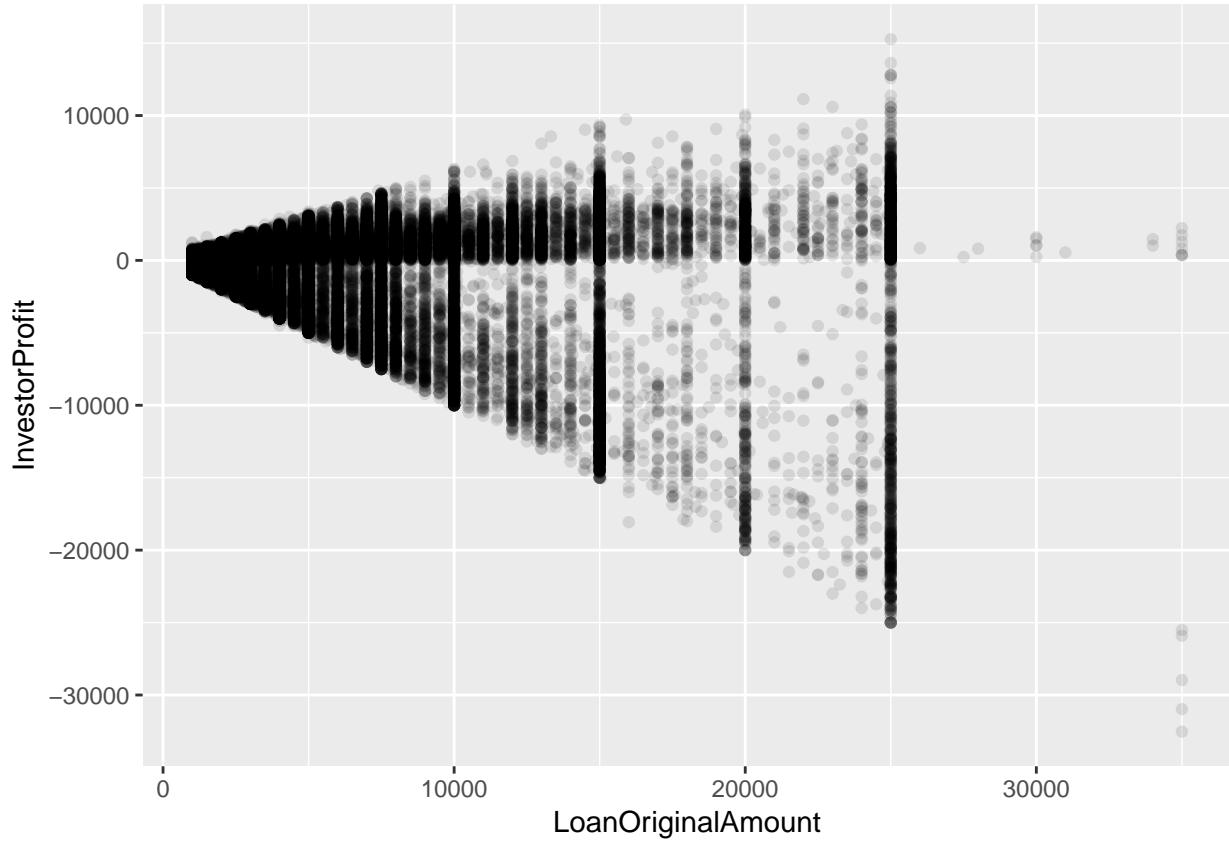
```
ggplot(data = subset(l1, ProsperRating..Alpha=="AA"), aes(x = BorrowerState)) + geom_bar(stat = "count")
  geom_text(stat = "count", aes(label=..count..), vjust = -0.5, size = 3)
```



Verilen kredi tutarına göre yatırımcıların kar ettiğleri tutarların karşılaştırılması aşağıda gösterilmiştir. Bu grafiğinde gösteriminde de ödemesi devam eden “*Current*” durumundaki hesaplar veri setinden çıkarılmıştır.

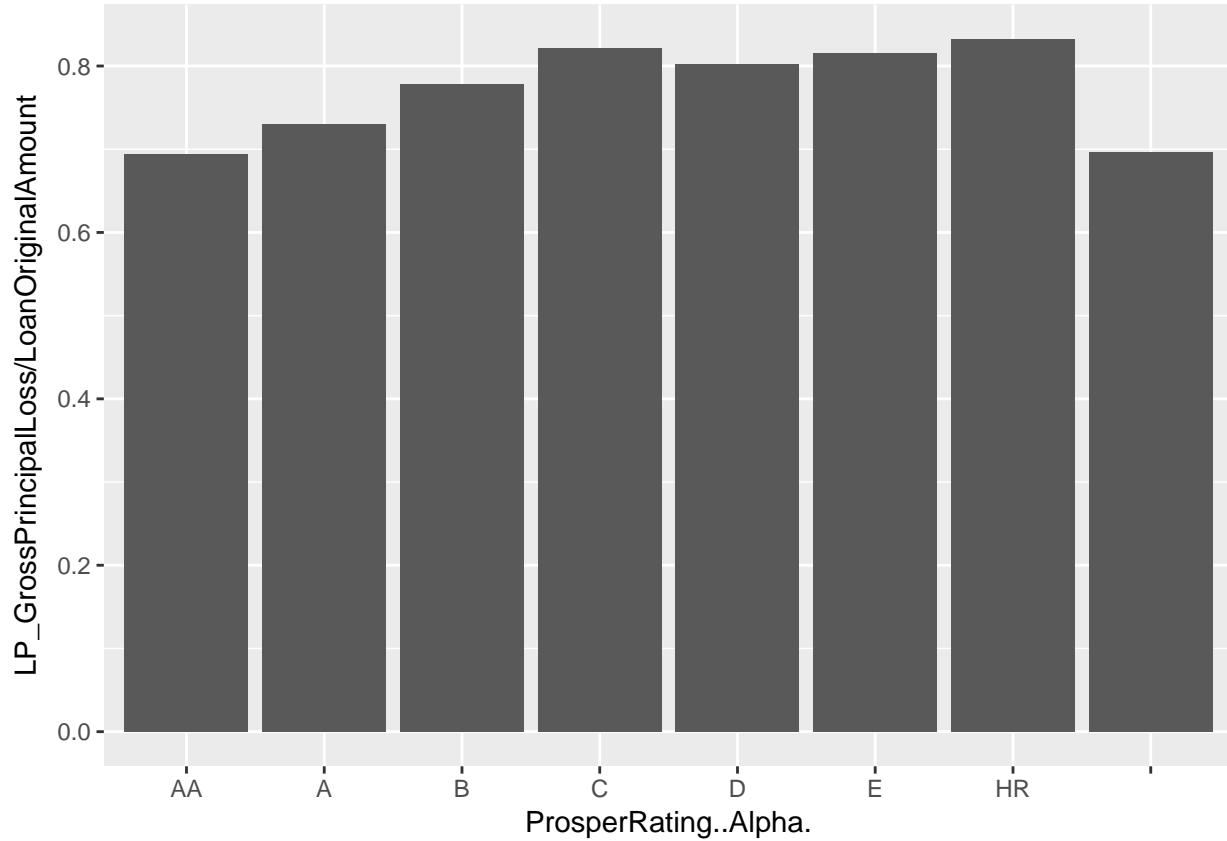
```
ld$InvestorProfit = ld$LP_CustomerPayments - ld$LoanOriginalAmount +
  ld$LP_ServiceFees + ld$LP_CollectionFees +
  ld$LP_NonPrincipalRecoverypayments
```

```
ggplot(data = subset(ld, LoanStatus != 'Current'), aes(x = LoanOriginalAmount, y = InvestorProfit)) + g
```



Borcunu ödemeyerek yatırımcıları zarara uğratan (*LP_GrossPrincipalLoss* değişkeni sıfırdan büyük bir değere sahip) hesaplarda, uğratılan zararın çekilen kredi tutarına oranında Prosper derecelendirmesine göre bir farklılık olup olmadığı yanıt aradığım bir diğer soru oldu. Aşağıdaki grafikte bu tür hesaplara ait brüt zarar tutarının (“*LP_GrossPrincipalLoss*”) çekilen kredi tutarına (“*LoanOriginalAmount*”) oranı y ekseninde gözlemlenirken, Prosper derecelendirmeleri x ekseninde temsil edilmektedir. Elde edilen grafiğe göre şu yorumu yapabiliyoruz; eğer çekilen kredi tutarının geri ödemesi yapılmıyorsa meydana gelen zarar tutarının çekilen kredi tutarına oranı tüm Prosper derecelendirmeleri için %70-%80 aralığındadır. Daha basit tabirle, eğer kişi borcunu ödememişse en iyi Prosper derecelendirmesine de sahip olsa borcunun ortalama yüzde yetmişini ödememiştir.

```
#Ödemesi yapılmayan hesaplarda Prosper derecelendirmesine göre; ödenmeyen tutarın, çekilen kredi tutarını
ggplot(data = subset(ld, LP_GrossPrincipalLoss > 0), aes(x = ProsperRating..Alpha., y = LP_GrossPrincipalLoss)) + geom_hex(bins = 100)
```

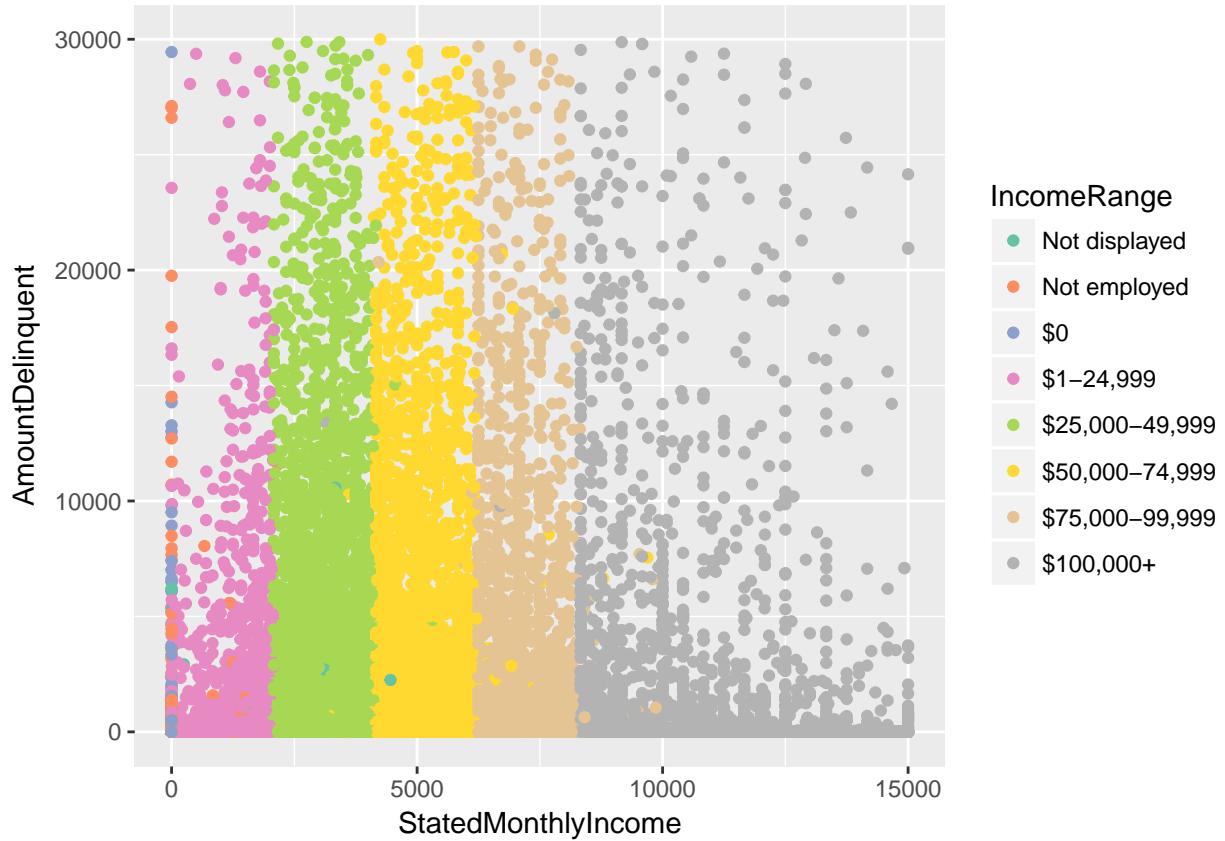


Çok değişkenli görseller

Gelir grupları farklı renklerle temsil edilmek üzere aylık gelirlerine (“*StatedMonthlyIncome*”) göre borçluların mevcut hesap ödenmemiş tutarları (“*AmountDelinquent*”) gösterilmiştir. Gelir gruplarının veri setindeki frekans dağılımlarına göre ödenmemiş tutarlar yoğunluk göstermektedir.

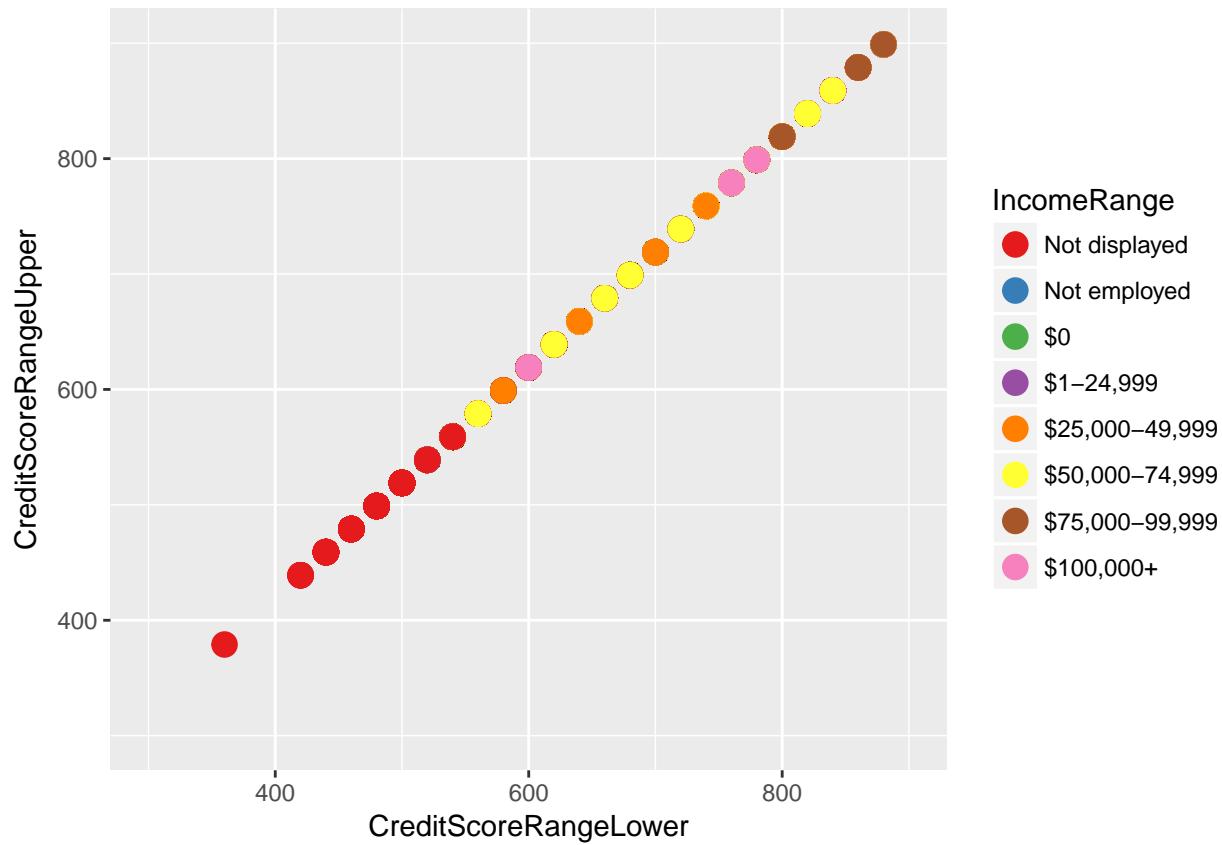
```
library(RColorBrewer)

ggplot(data = subset(ld, !is.na(IncomeRange)), aes(x = StatedMonthlyIncome, y = AmountDelinquent, color
  geom_point() + xlim(0,15000) + ylim(0,30000) +
  scale_color_brewer(palette="Set2")
```

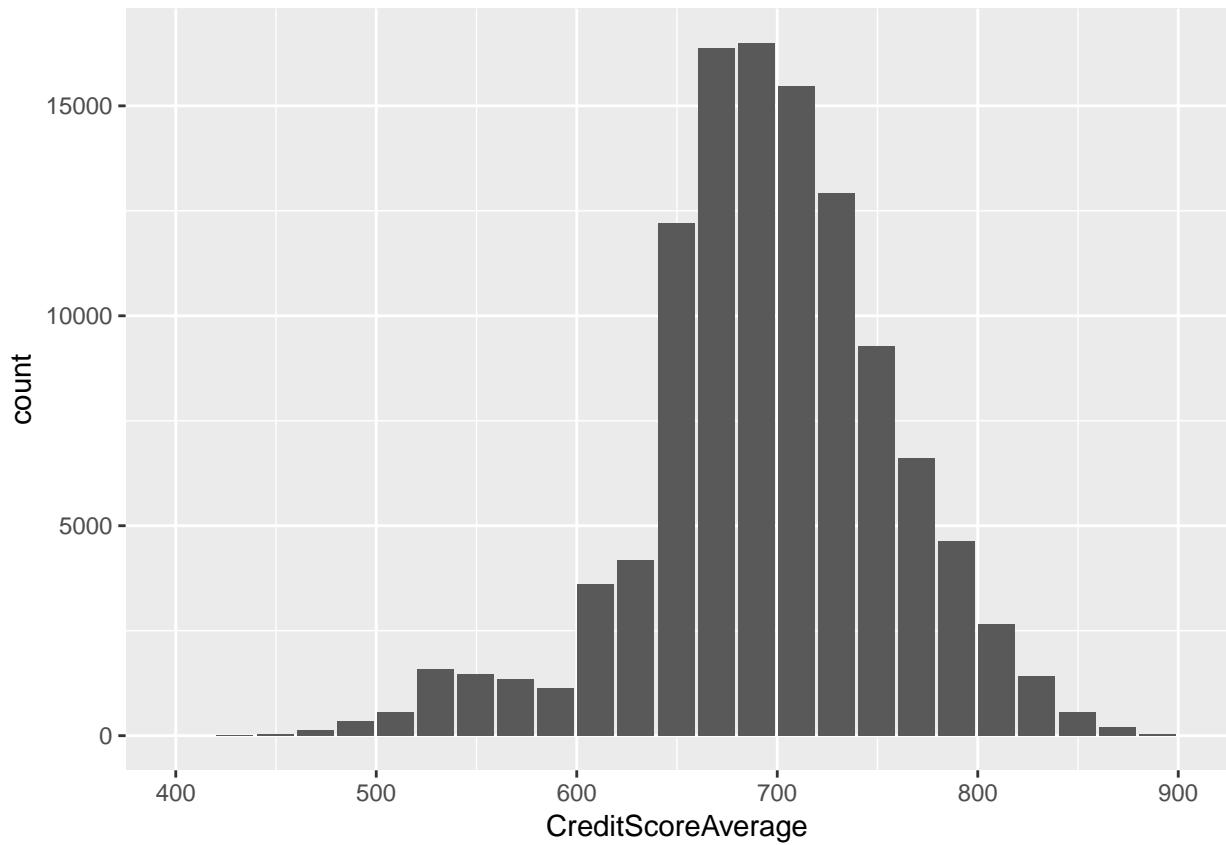


Veri setinde her gözleme ait alt ve üst değerler olmak üzere iki kredi skoru bulunmaktadır, buna göre gelir gruplarının kredi skorlarına göre dağılımı gösterilmiştir. Grafikten görüldüğü üzere dahil olunan gelir grubunun kredi skoruyla bir ilişkisi bulunmamaktadır. Örneğin, senelik yüz bin dolar üzeri kazanan bir borçlunun 50-75 bin aralığına dahil bir borçludan kredi skoru düşük olabilmektedir. Sonraki adımda, "CreditScoreAverage" adlı değişken oluşturularak alt ve üst kredi skor değerlerinin ortalamasının alınmasıyla kredi skoru tek değere indirgemistiştir. "CreditScoreAverage" a göre borçların normal dağılım yatığı, kredi skorlarının çoğunuğunun 650-750 aralığında toplandığı gözlemlenmektedir.

```
ggplot(data = ld, aes(x = CreditScoreRangeLower, y = CreditScoreRangeUpper, colour = IncomeRange)) +
  geom_point(size=4) + xlim(300,900) + ylim(300,900) +
  scale_color_brewer(palette="Set1")
```

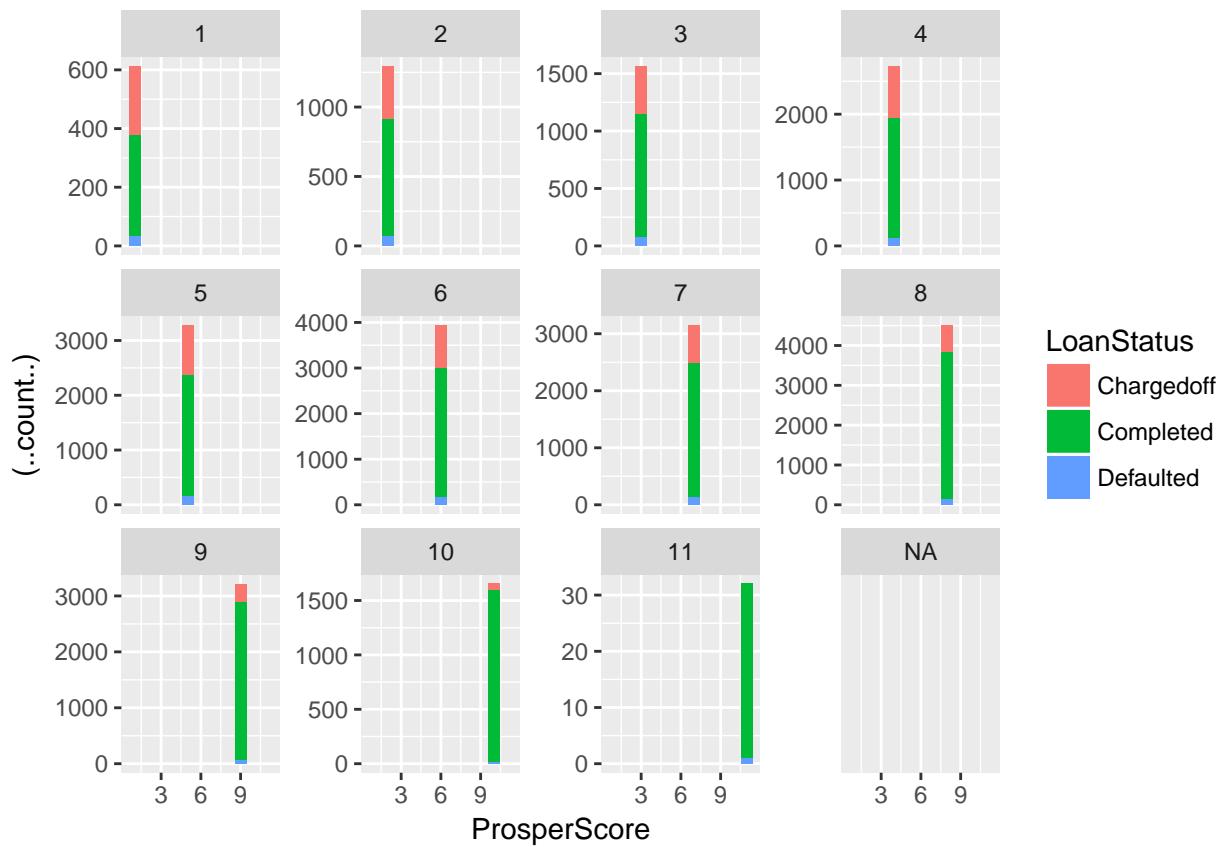


```
ld$CreditScoreAverage <- (ld$CreditScoreRangeLower + ld$CreditScoreRangeUpper)/2
ggplot(data = subset(ld, !is.na(CreditScoreAverage)), aes(x = CreditScoreAverage)) + geom_bar() + xlim(
```



Her bir Prosper derecelendirmesi için “chargedoff”, ödemesi yapılmamış(“Defaulted”) ve ödemesi tamamlanmış (“Completed”) hesap durumlarının dağılımı gösterilmiştir. Elde edilen sonuca göre Prosper derecesi arttıkça her bir derecelendirme değeri içinde ödemesi tamamlanmış (“Completed”) hesapların sayısı artmaktadır.

```
ggplot(data = subset(l1, LoanStatus == 'Completed' | LoanStatus == 'Defaulted' | LoanStatus == 'Chargedoff')) +  
  geom_bar(aes(y = (.count...))) +  
  facet_wrap(~ProsperScore, scales = "free_y")
```



Prosper.com'un faaliyete geçtiğinden bu yana oluşturulan kredi tutarları ve açılan hesap sayısının zamana göre değişimi platformun başarımını izleme açısından önemli veriler olarak incelenmiştir. Çıktılar, çeyrek-yıllara göre zaman serisi grafiği üzerrinde gösterilmiştir. Bu noktada, karşılaşduğum engelden dolayı küçük bir düzenleme yaptım. Veri setinde kredi hesabı oluşturma tarihine göre yıl ve çeyrek sayısını gösteren "LoanOriginationQuarter" değişkeninde format yıl-çeyrek-saat-dakika-saniye şeklinde olduğundan dolayı grafik üzerinde zaman serisinin gösterildiği x ekseninde sıralama önce birinci çeyrek için yıllar ardışık olarak sıralanıyor, sonra ikinci çeyrek için tekrar yıllar başlangıç yılından itibaren ardışık olarak sıralanıyor ve dördüncü çeyreğe kadar olan bu sıralama okuyucu için istenen grafiği vermeyen bir formattaydı. Bunun için `separate()` ve `unite()` fonksiyonları kullanılarak "LoanOriginationQuarter" değişkeninde önce çeyrek sonra yıl ifadesinin yazılması sağlanmış, tamamı 00:00:00 şeklinde olmasından ötürü gereksiz olan saat, dakika, saniye belirten kısım atılmıştır.

#Zaman serisi grafiğinde yılların çeyreklerde göre sıralı olması için düzenleme:

```
ld$LoanOriginationQuarter <- as.character(ld$LoanOriginationQuarter)
ld <- separate(ld, LoanOriginationQuarter, c("Quarters", "Year"), sep = " ")
ld <- unite(ld, col = LoanOriginationQuarter, Year, Quarters, sep = " ")
```

Yapılan düzenleme sonrası çeyrek-yıllara göre oluşturulan kredi tutarı grafiği elde edilmiştir. Grafikte dikkat çeken durum ise 2009 yılı başlangıcına ait faaliyetin olmayışı. Bunun sebebiyse Prosper.com'un burada kısaca ABD yetkili makamlarıyla arasındaki bürokratik işlemler nedeniyle diye açıklayabileceğim sessiz döneme girmesi. İlgili düzenlemelerin yapılması sonrasında oluşturulan kredi tutarlarında yükselerek artan bir grafik görülmektedir.

```
Origination_QY <- ld %>%
  select(Quarter = LoanOriginationQuarter, Amount = LoanOriginalAmount) %>%
  group_by(Quarter) %>%
  summarise(Loans = n()/ 10 ^ 3, Dollars = sum(Amount)/ 10 ^ 6) %>%
```

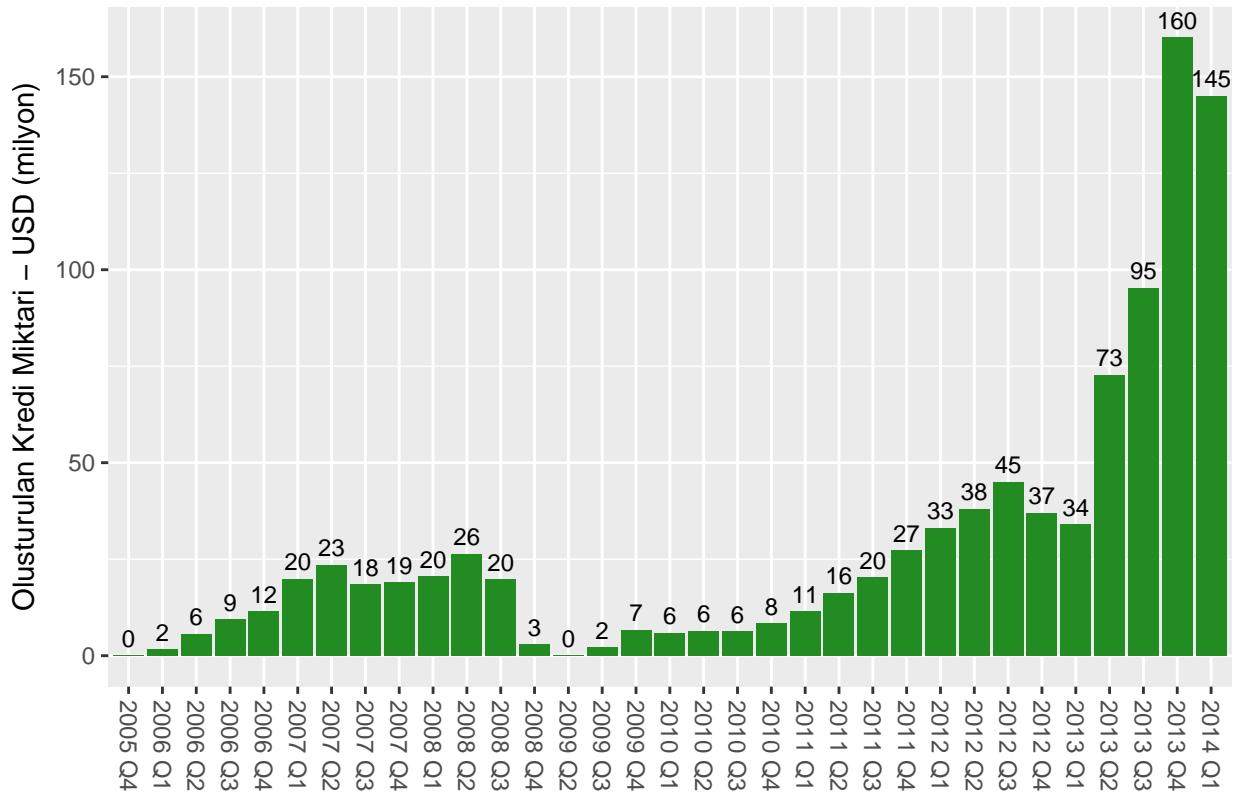
```

arrange(Quarter) %>%
filter(Quarter < "2014 Q3")

ggplot(Origination_QY, aes(x = Quarter, y = Dollars)) +
  geom_bar(stat = "identity", fill = "forestgreen") +
  geom_text(aes(label = round(Dollars, 0)), vjust = -0.5, size = 3) +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5),
        axis.title.x = element_blank()) +
  ylab("Oluşturulan Kredi Miktarı - USD (milyon)") +
  ggtitle("Çeyrek-Yıllara göre oluşturulan kredi tutarı (USD)")

```

Çeyrek-Yıllara göre oluşturulan kredi tutarı (USD)



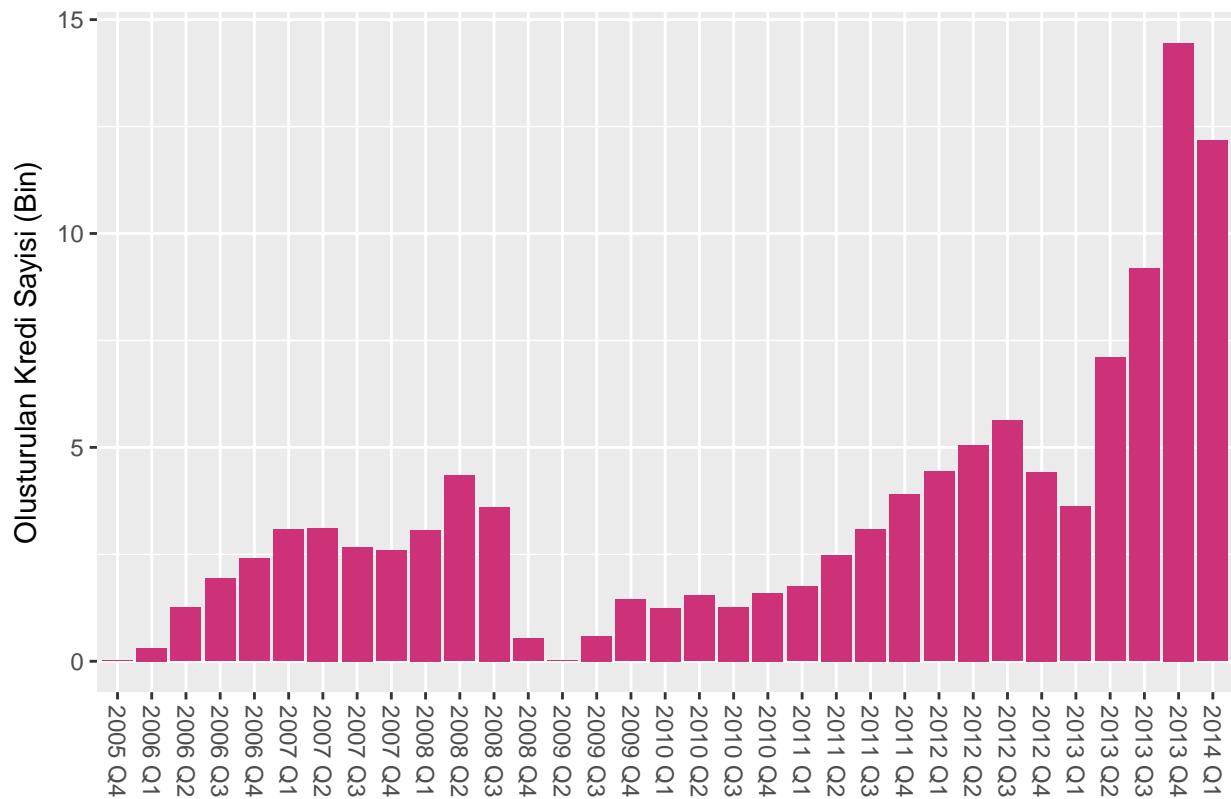
Oluşturulan kredi sayısı da kredi tutarları grafiğiyle uyusan, benzer bir grafik vermektedir.

```

ggplot(Origination_QY, aes(x = Quarter, y = Loans)) +
  geom_bar(stat = "identity", fill = "violetred3") +
  theme(axis.text.x = element_text(angle = -90, vjust = 0.5),
        axis.title.x = element_blank()) +
  ylab("Oluşturulan Kredi Sayısı (Bin)") +
  ggtitle("Çeyrek-yıllara göre oluşturulan kredi sayısı")

```

Çeyrek–yillara göre oluşturulan kredi sayısı



Korelasyonlar

Seçmiş olduğum çeşitli nümerik değişkenler arasında korelasyon olup olmadığını `cor.test()` fonksiyonu ile gözlemedim. (Yorum satırında korelasyon tablosu oluşturmayı sağlayan kodlar daha sonra kullanılmak üzere bırakılmıştır.)

```
cor.test(1d$CreditScoreAverage, 1d$StatedMonthlyIncome)
```

```
##
## Pearson's product-moment correlation
##
## data: 1d$CreditScoreAverage and 1d$StatedMonthlyIncome
## t = 36.54, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1021433 0.1136511
## sample estimates:
##      cor
## 0.1079008
```

```
cor.test(1d$CreditScoreAverage, 1d$LoanOriginalAmount)
```

```
##
## Pearson's product-moment correlation
##
## data: 1d$CreditScoreAverage and 1d$LoanOriginalAmount
```

```

## t = 122.07, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.3357190 0.3460095
## sample estimates:
## cor
## 0.3408745

#corlist <- function(df, column) {
#  m <- c()
#  for (col in names(df)) {
#    foo <- tryCatch(round(cor(as.numeric(df[, column]),
#                               as.numeric(df[, col])), 2),
#                    error = function(e)NA)
#    m <- append(m, foo)
#  }
#  names(m) <- names(df) # assign row names
#  m
#}

## calculate correlations between all values in a dataframe
## (could not use cor(dataframe) because it does not coerce to numeric)
#cortable <- function(df){
#  m <- c()
#  for (col in names(df)) {
#    m <- cbind(m, suppressWarnings(corlist(df, col)))
#  }
#  m <- data.frame(m)
#  names(m) <- names(df) # assign column names
#  m
#}
#cortable(ld)

```

“GoodmanKruskal” ile kategorik değişkenler ile nümerik değişkenler arasındaki korelasyonu inceledim. Elde edilen sonuçlar çok düşük değerler olduğu için araştırdığım değişkenler arasında bir korelasyon bulunamamıştır. Ayrıca korelasyon konusu istatistik disiplini içinde başı başına ayrı bir konu olduğu için bu paylaşım içinde detaya girmemiştir.

```

library(GoodmanKruskal)
#https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html
GKtau(ld$ProsperRating..Alpha., ld$Occupation)

##           xName          yName Nx Ny tauxy tauyx
## 1 ld$ProsperRating..Alpha. ld$Occupation 8 68 0.002 0.014
GKtau(ld$ProsperScore, ld$IsBorrowerHomeowner)

##           xName          yName Nx Ny tauxy tauyx
## 1 ld$ProsperScore ld$IsBorrowerHomeowner 12 2 0.012 0.002

```