MA684 Midterm Project

Dongyuan Zhou

November 13, 2017

Background

In the financial field, making decisions on whether to lend money to borrowers is one of the most important work. A whole cycle of this work usually includes the following two steps.

Firstly, grades the loan according to the borrower credit record as well as the loan amount and loan time, and then make decision whether lend or not.

Secondly, summarize the default rate and improve the former step to avoid default next time.

Objective

In this analysis, we analyze data from Lending club (LC):

Stage1: Initial EDA (Data: 2015-2017)

Analyze how borrower's status inluenced their loan amount? (Linear regression)

Stage2: Before loan was funded (Data: 2015-2017)

Analyze how LC assigned loan grade: How borrower's status as well as the loan amount and loan time influenced loan grade? (Multinomial regreesion)

Stage3: After laon ended (Data: 2007-2011)

Summarize default rate for each state.(Multilevel logistic regression)

Stage4: Summarize and Discussion

Assessment of the result. Discuss about the data limitations and future directions.

Data description

Data source: *https://www.lendingclub.com/info/download-data.action*

To get reasonable analysis, we only choose the data which had been verified by LC.

In the whole analysis, we transform loan grade to grade number.

grade	A	B	C	D	E	F	G
gradenumber	7	6	5	4	3	2	1

Code for data cleaning: "data clean.R"

```
## Final Data
Loan1517 <- read.table("Loan1517.csv", header = TRUE, sep = ",")
Loan0711 <- read.table("Loan0711.csv", header = TRUE, sep = ",")
Loan1517$year <- ifelse(Loan1517$term == "36",3,5)
Loan0711$year <- ifelse(Loan0711$term == "36",3,5)
Loan0711$default <- ifelse(Loan0711$loan_status == "Fully Paid",0,1)
Loan1517$region <- tolower(Loan1517$region)
```

```
Loan0711$region <- tolower(Loan0711$region)
Loan0711$dti <- Loan0711$dti/100
Loan1517$dti <- Loan1517$dti/100
```

1. Initial EDA

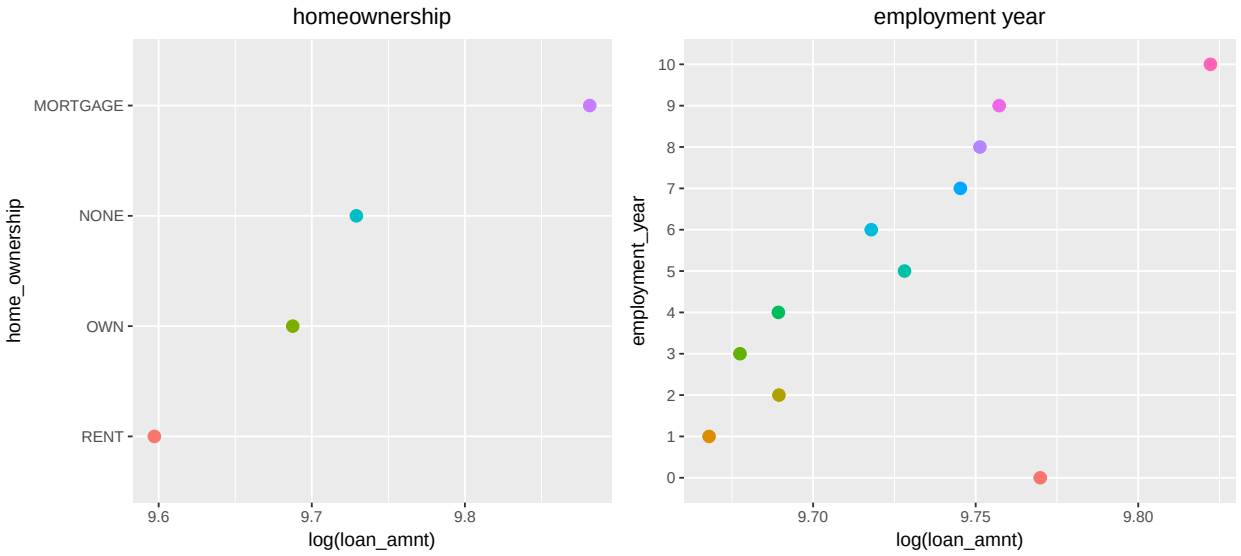
Analyze how borrower's status inluenced their loan amount? (Linear regression)

```
# 1. How home ownership influenced loan amount?
home <- summarise(group_by(Loan1517,Loan1517$home_ownership),
                  loan_amnt = mean(loan_amnt, na.rm = TRUE))
home$home_ownership <- home$`Loan1517$home_ownership`
p1 <- ggplot(home)+
  aes(x=reorder(home$home_ownership,loan_amnt),y = log(loan_amnt),
      color=reorder(home$home_ownership,loan_amnt))+
  geom_point(size=3)+
  xlab("home_ownership")+
  labs(title = "homeownership")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")+
  coord_flip()

# 2. How employment years inluenced loan amount?
emp_year <- summarise(group_by(Loan1517,Loan1517$emp_year),
                      loan_amnt = mean(loan_amnt, na.rm = TRUE))
emp_year$emp_years <- emp_year$`Loan1517$emp_year`
p2 <- ggplot(emp_year)+
  aes(factor(emp_years), y =  log(loan_amnt),color = factor(emp_years))+
  geom_point(size=3)+
  xlab("employment_year")+
  labs(title = "employment year")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")+
  coord_flip()

# 3. How loan purpose influenced loan amount?
loanpurpose <- summarise(group_by(Loan1517,Loan1517$purpose),
                         loan_amnt = mean(loan_amnt, na.rm = TRUE))
loanpurpose$loan_purpose <- loanpurpose$`Loan1517$purpose`
p5 <- ggplot(loanpurpose)+
  aes(reorder(loanpurpose$loan_purpose,loan_amnt),y = log(loan_amnt),
      color=reorder(loanpurpose$loan_purpose,loan_amnt))+
  geom_point(size=3)+
  xlab("purpose")+
  labs(title = "purpose")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")+
  coord_flip()

# 4. How state status influenced loan amount?
us <- map_data("state")
arr <- USArrests %>%
  add_rownames("region") %>%
  mutate(region=tolower(region))
stateofloan <- Loan1517%>%
  group_by(region)%>%
```

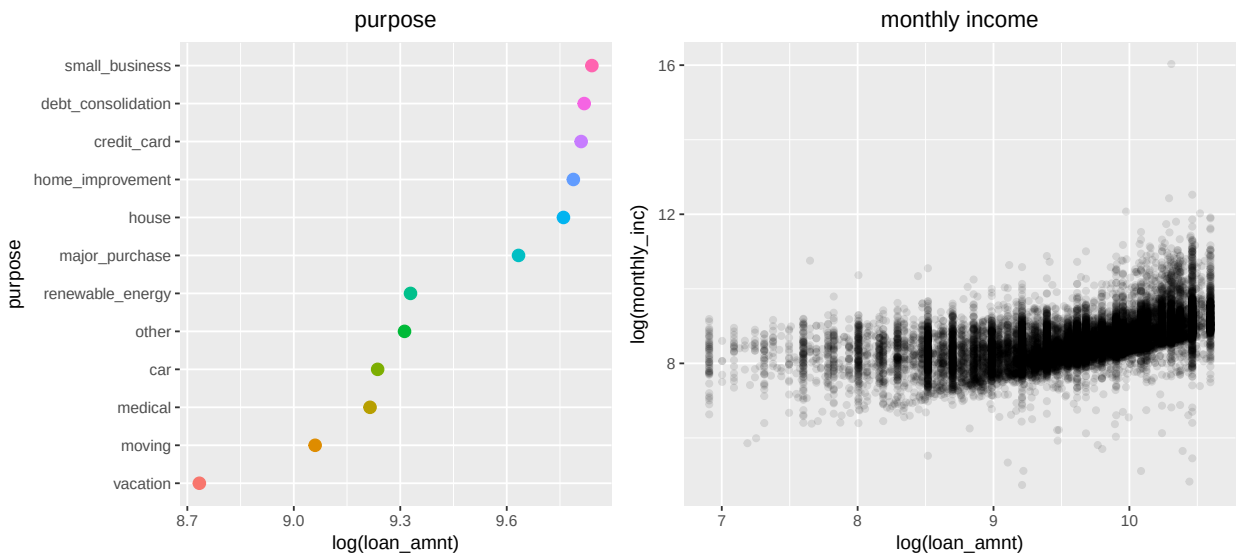
```
  summarise(amount = sum(loan_amnt), abbr = unique(addr_state))
stateofloan <- merge(arr, stateofloan,by = "region")
p3 <- ggplot()+
  geom_map(data=us, map=us, aes(x=long, y=lat, map_id=region),
           fill="#ffffff", color="#ffffff", size=0.15)+
  geom_map(data=stateofloan, map=us, aes(fill=log(amount), map_id=region),
           color="#ffffff", size=0.15)+
  scale_fill_continuous(low='thistle2', high='darkred',guide='colorbar')+
  labs(x=NULL, y=NULL,title = "state")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")
  coord_map("albers", lat0 = 39, lat1 = 45)
```

```
## <ggproto object: Class CoordMap, Coord>
##     aspect: function
##     distance: function
##     is_linear: function
##     labels: function
##     limits: list
##     orientation: NULL
##     params: list
##     projection: albers
##     range: function
##     render_axis_h: function
##     render_axis_v: function
##     render_bg: function
##     render_fg: function
##     train: function
##     transform: function
##     super:  <ggproto object: Class CoordMap, Coord>
```

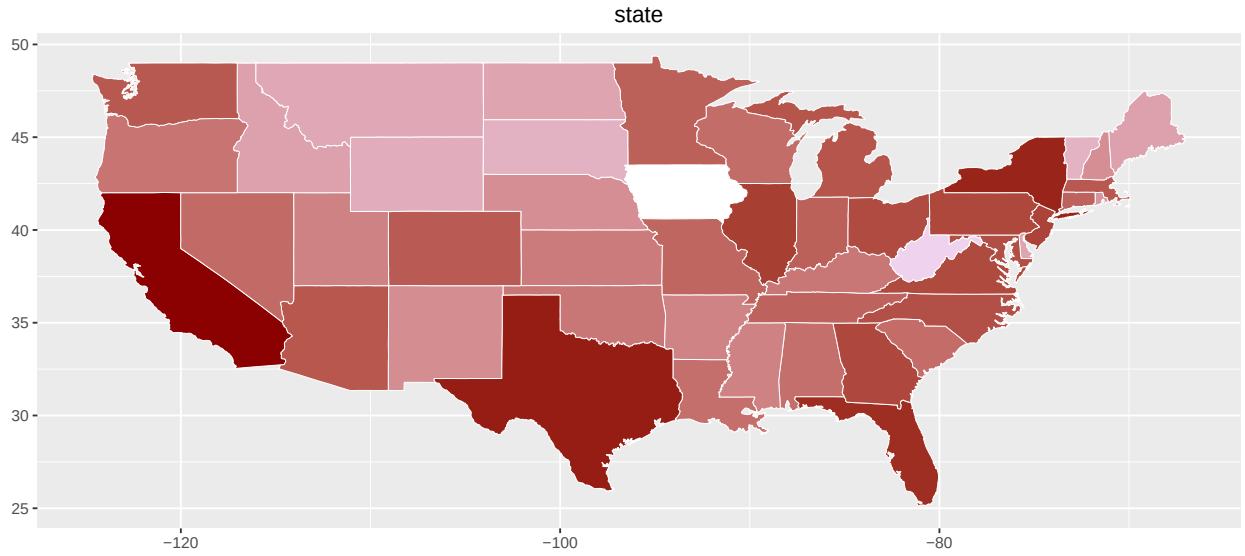
```
# no data for Iowa

# 5. How monthly income influenced loan amount?
income <- summarise(group_by(Loan1517,Loan1517$monthly_inc),
                    loan_amnt = mean(loan_amnt, na.rm = TRUE))
income$monthly_inc <- income$`Loan1517$monthly_inc`
p4 <- ggplot(income)+
  aes(x = log(monthly_inc), y =  log(loan_amnt))+
  geom_point(alpha=0.1)+
  coord_flip()+
  labs(title = "monthly income")+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(p1,p2,ncol =2)
```




```
grid.arrange(p5,p4,ncol =2)
```

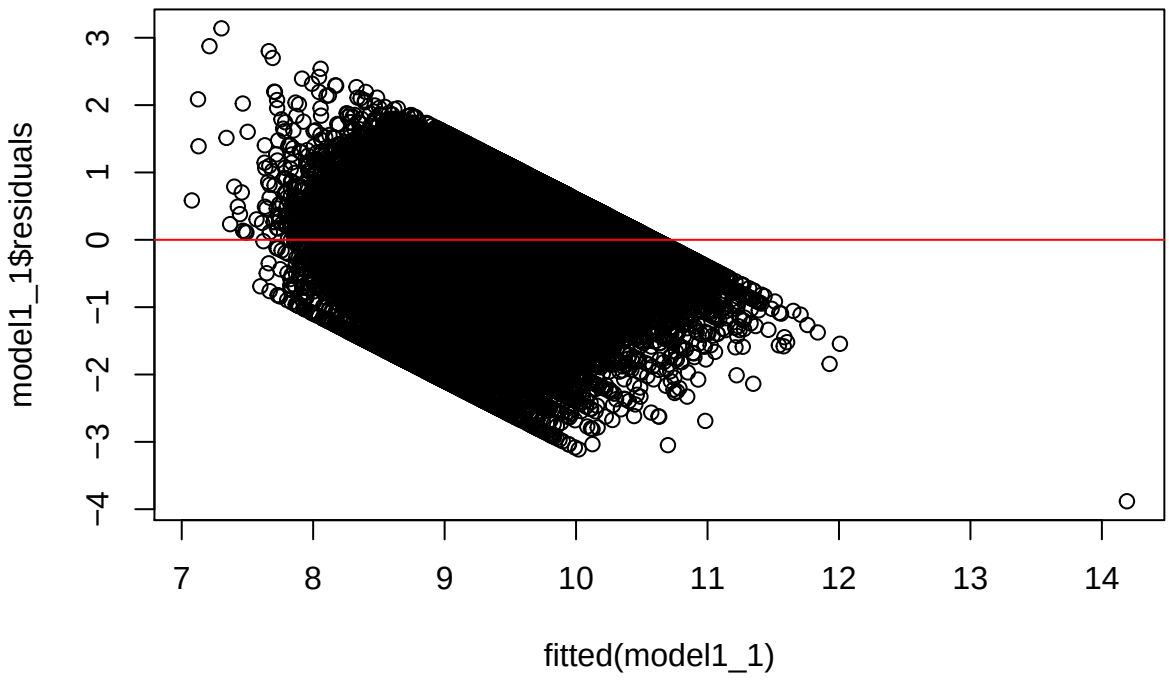



```
p3
```



From the chart above, we could see that borrowers whose home is mortgage had greater loam amount than the other three types. For the employment length, it seems that borrowers who was just get their jobs had more loan amount than those who had already work for years. Borrowers who loaned money for business, they had greater loan amount than those who loaned just for vacation. For the monthly income, borrowers who receive higher income per month could have higher loan amount, it is reasonable since they seemd to have higher debt-paying ability, and they seemed to have more source of demand for money. From the map we could clear see that loan amount in CA, TX, FL and NY is higher than other states, which reflects the financial demand in each state.

```
# fit model
model1_1 <- lm(log(loan_amnt) ~ log(monthly_inc) + purpose + home_ownership +
                emp_year + addr_state, data = Loan1517)
plot(fitted(model1_1),model1_1$residuals)
abline(0,0, col="red")
```



Check the residual: Residual standard error: 0.6091

Interpret:

For each 1% difference in monthly income, the predicted difference in loan amount is 0.63%.

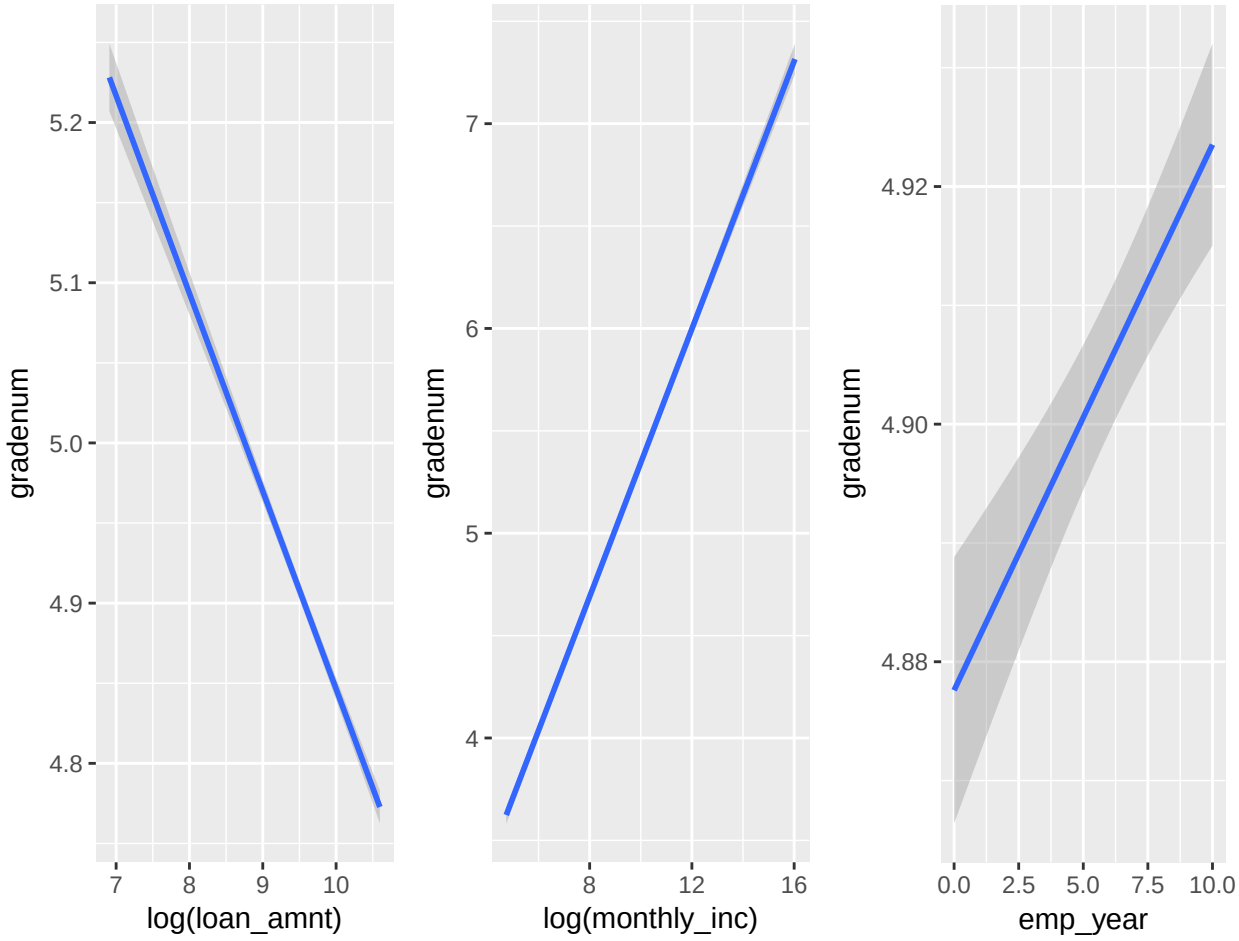
For each 1% difference in employment year, the predicted difference in loan amount is 0.003%.

The other input, purpose, home ownership, state, is categorical so it does not make sense to take its logarithm.

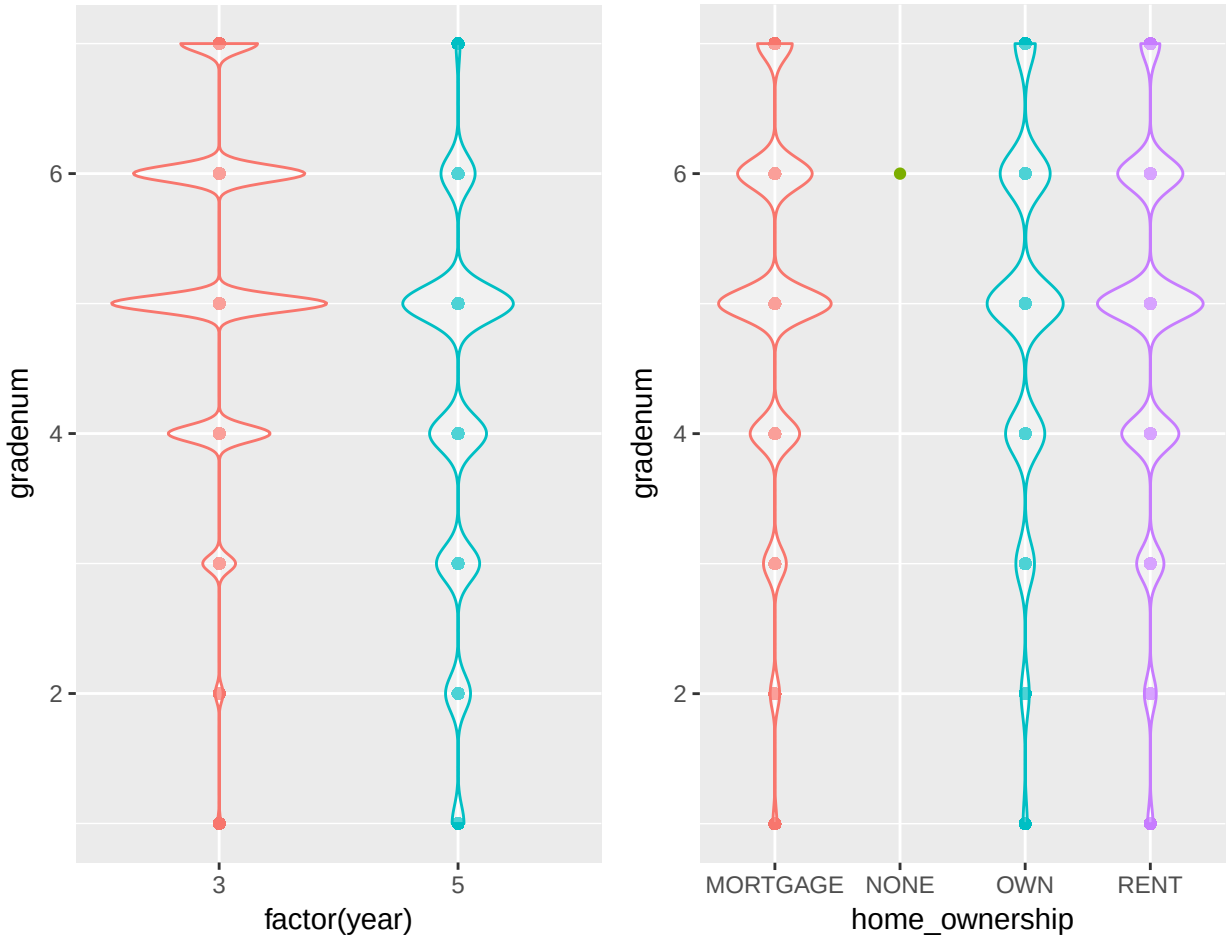
2. Before loan was funded (Data:2015-2017)

Analyze how LC assigned loan grade: How borrower's status as well as the loan amount and loan time influenced loan grade? (Multinomial regreesion)

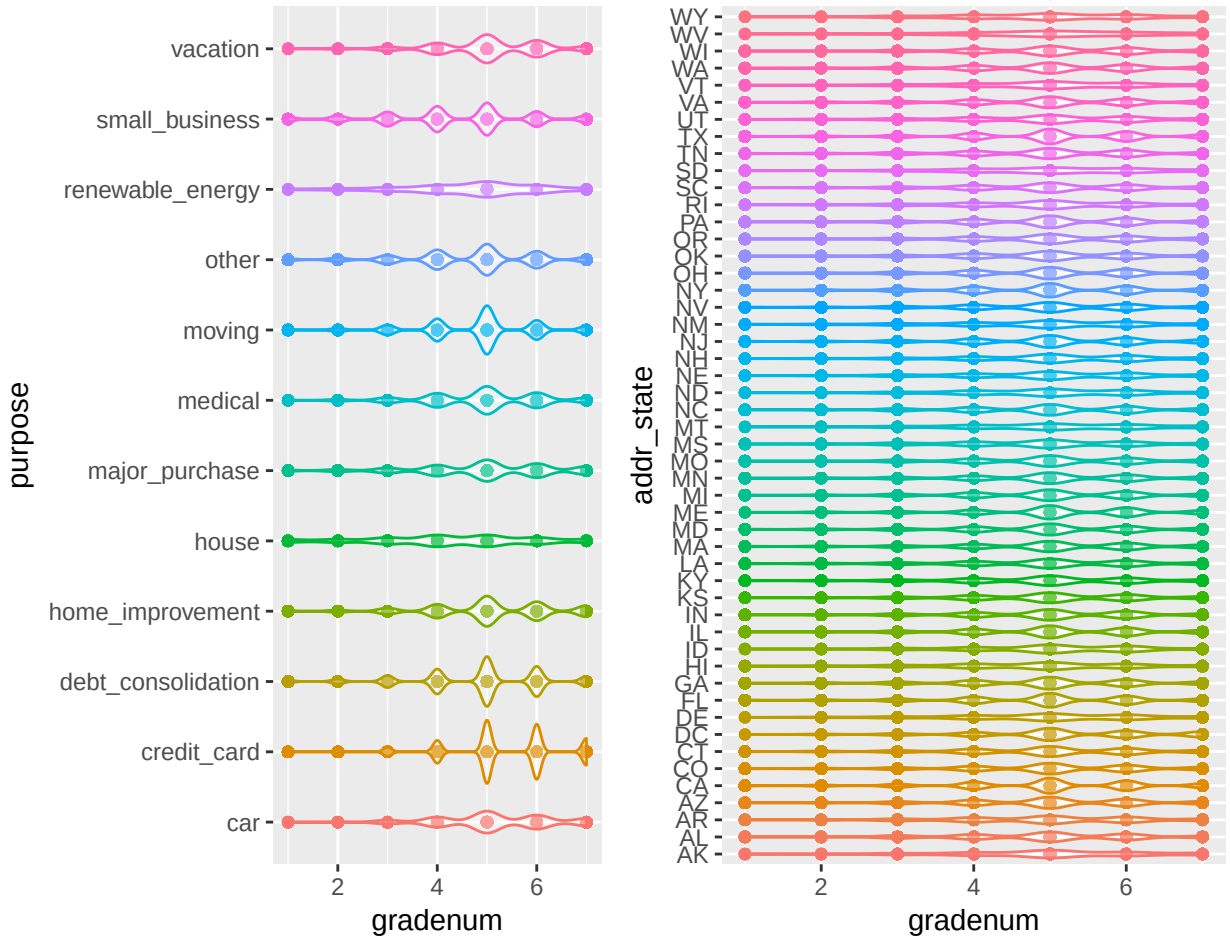
```
## intial EDA
grid.arrange(
ggplot(Loan1517)+
  aes(x=log(loan_amnt),y=gradenum)+
  geom_smooth(method="lm"),
ggplot(Loan1517)+
  aes(x=log(monthly_inc),y=gradenum)+
  geom_smooth(method="lm"),
ggplot(Loan1517)+
  aes(x=emp_year,y=gradenum)+
  geom_smooth(method="lm"),
ncol=3
)
```




```
grid.arrange(
ggplot(Loan1517)+
  aes(x=factor(year),y=gradenum,color=factor(year))+
  geom_point()+
  geom_violin(alpha=0.3)+
  theme(legend.position="none"),
ggplot(Loan1517)+
  aes(x=home_ownership,y=gradenum,color=home_ownership)+
  geom_point()+
  geom_violin(alpha=0.3)+
  theme(legend.position="none"),
ncol=2
)
```




```
grid.arrange(
ggplot(Loan1517)+
  aes(x=purpose,y=gradenum,color=purpose)+
  geom_point()+
  geom_violin(alpha=0.3)+
  theme(legend.position="none")+
  coord_flip(),
ggplot(Loan1517)+
  aes(x=addr_state,y=gradenum,color=addr_state)+
  geom_point()+
  geom_violin(alpha=0.3)+
  theme(legend.position="none")+
  coord_flip(),
ncol=2
)
```



From the chart above, we could see that higher loan amount, lower loan grade; lower income, lower loan grade; What's more, longer loan term means higher probability to have lower loan grade and borrowers who rent their home seems to has lower loan grade. There is not much significant diffenrence according the the EDA for loan purose and loan state. Maybe we could use the multilevel regression to show the result.

```
## fit models
## model2_0
model2_0<- polr(grade ~ log(loan_amnt) + year + log(monthly_inc) +
                  home_ownership + emp_year + purpose + addr_state,
                data = Loan1517)
```

Check residuals: Residual Deviance: 567279.20

AIC: 567425.20

Interpret:

statistical significant: Loan amount, term, monthly income

not significant: addrstate, employment year, purpose,home ownership

We get emp_year that is positive and insignificant contrary to our expectation. It seems reasonable to remove emp_year variable from our model. Meanwhile, according to the industry regulation, loan purpose and state of loan has little significance compared to monthly income and loan amount.

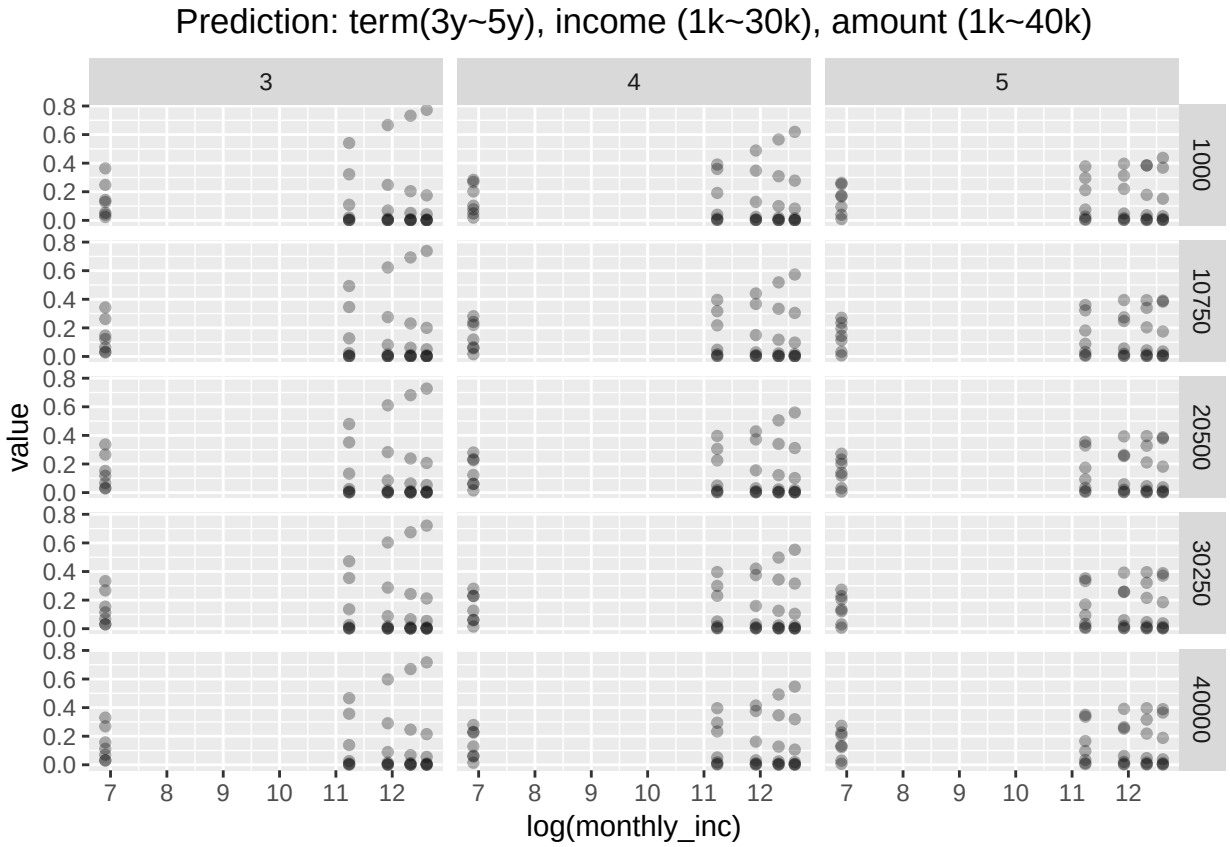
Therefore, we consider a new model for the grade of loan as a function of loan amount, loan term and borrower's monthly income.

```
## model2_1
model2_1<- polr(grade ~ log(loan_amnt) + year + log(monthly_inc),
                data = Loan1517)
summary(model2_1)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = grade ~ log(loan_amnt) + year + log(monthly_inc),
##     data = Loan1517)
##
## Coefficients:
##                      Value Std. Error t value
## log(loan_amnt)    0.08061   0.006914   11.66
## year              0.73920   0.005199  142.18
## log(monthly_inc) -0.77201   0.008807  -87.66
##
## Intercepts:
##     Value    Std. Error t value
## A|B  -5.7369   0.0705   -81.3740
## B|C  -4.0606   0.0696   -58.3036
## C|D  -2.3739   0.0693   -34.2518
## D|E  -1.2098   0.0693   -17.4471
## E|F  -0.0846   0.0698    -1.2120
## F|G   1.1674   0.0714    16.3443
##
## Residual Deviance: 575377.17
## AIC: 575395.17
```

```
predx<-expand.grid(loan_amnt=seq(1000,40000,length.out=5),
                   year=3:5,monthly_inc=seq(1000,300000,length.out=5))
predy<-predict(model2_1,newdata=predx,type="prob")

ggplot(melt(data.frame(predx,predy),
            id.vars = c("loan_amnt", "year", "monthly_inc")))+
  geom_point(alpha=0.3)+
  aes(x=log(monthly_inc),y=value,group=variable)+
  facet_grid(loan_amnt~year,scale="free_x")+
  labs(title = "Prediction: term(3y~5y), income (1k~30k), amount (1k~40k)")+
  theme(plot.title = element_text(hjust = 0.5))
```



Check rediduals: Residual Deviance: 575377.17

AIC: 575395.17

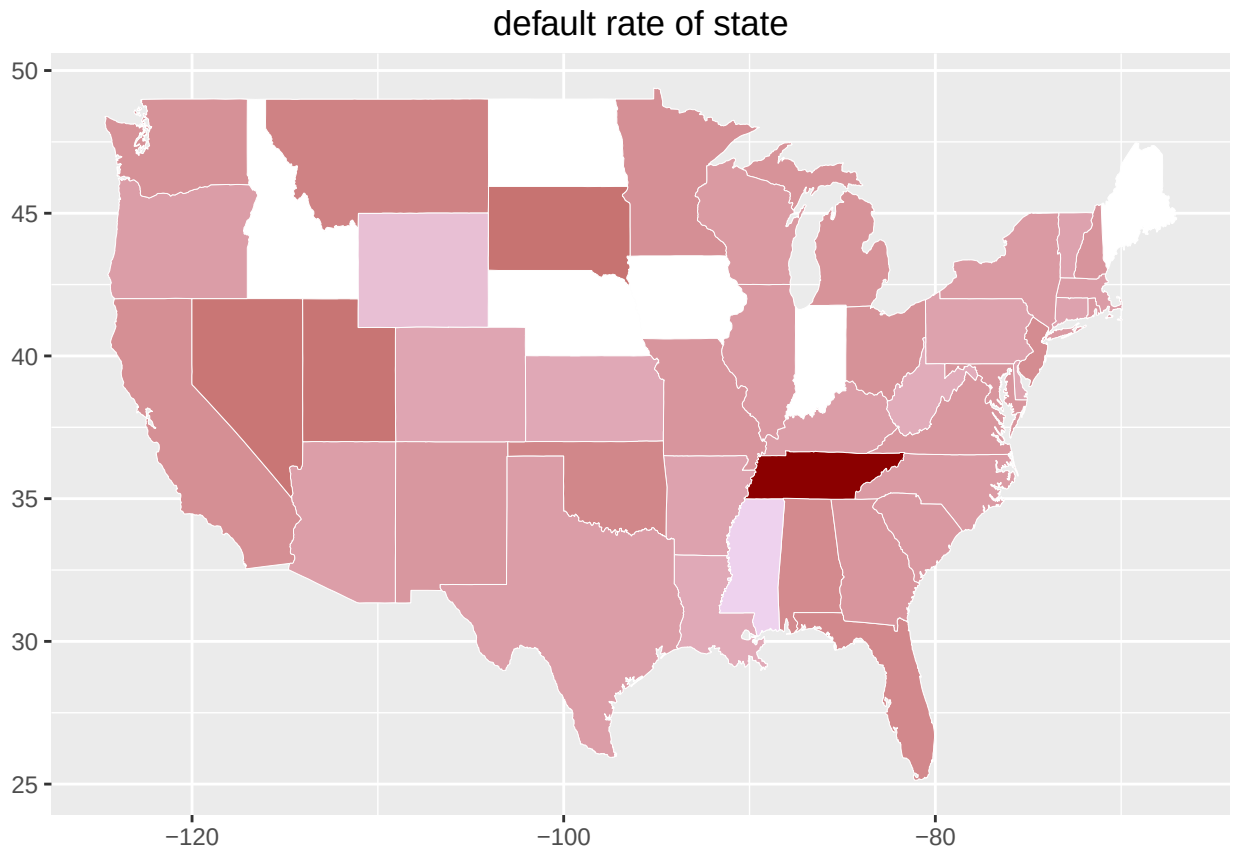
3. After laon ended (Data: 2007-2011)

Summarize default rate for each state.(Multilevel logistic regression)

dti: Debt to income ratio

```
## fixed effect
default <- Loan0711%>%
  group_by(region)%>%
  summarise(defrate = sum(default)/length(default))
default <- merge(arr, default,by = "region")
ggplot()+
  geom_map(data=us, map=us, aes(x=long, y=lat, map_id=region),
           fill="#ffffff", color="#ffffff", size=0.15)+
  geom_map(data=default, map=us, aes(fill=defrate, map_id=region),
           color="#ffffff", size=0.15)+
  scale_fill_continuous(low='thistle2', high='darkred',guide='colorbar')+
  labs(x=NULL, y=NULL,title = "default rate of state")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")
```

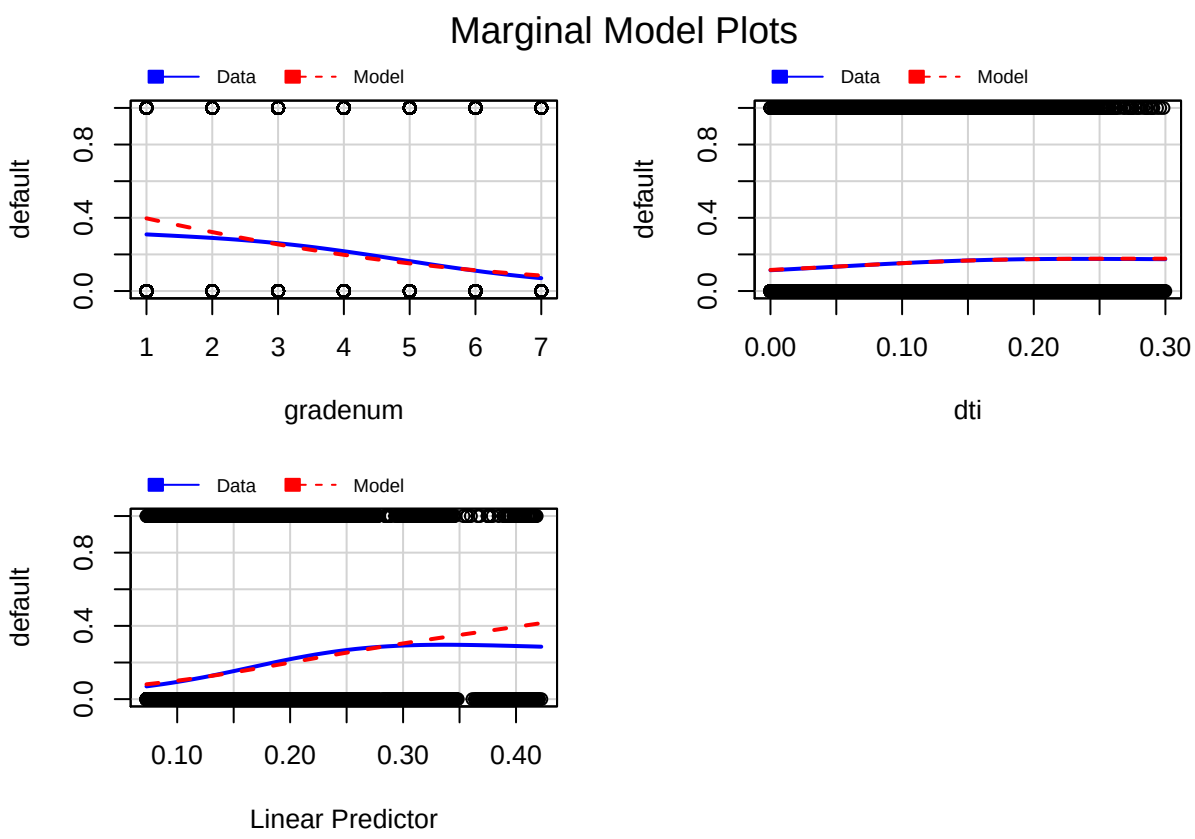
```
## Warning: Ignoring unknown aesthetics: x, y
```



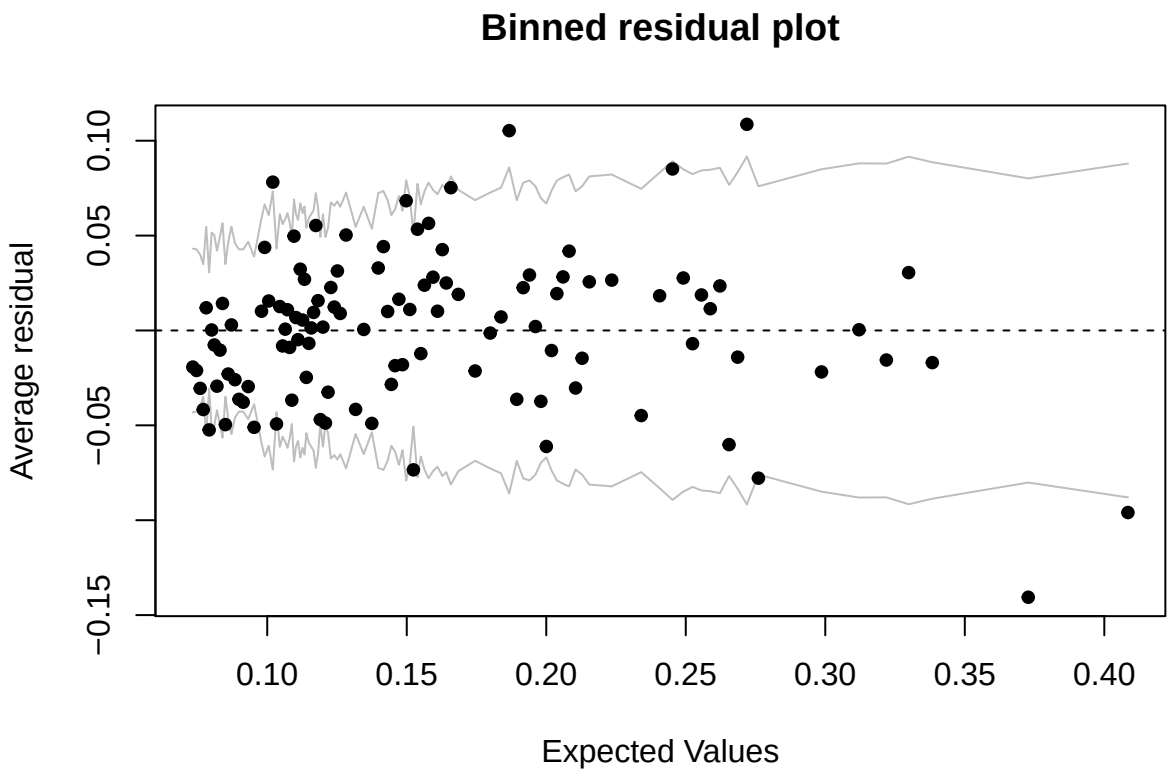

```
  coord_map("albers", lat0 = 39, lat1 = 45)
```

```
## <ggproto object: Class CoordMap, Coord>
##     aspect: function
##     distance: function
##     is_linear: function
##     labels: function
##     limits: list
##     orientation: NULL
##     params: list
##     projection: albers
##     range: function
##     render_axis_h: function
##     render_axis_v: function
##     render_bg: function
##     render_fg: function
##     train: function
##     transform: function
##     super:  <ggproto object: Class CoordMap, Coord>
```

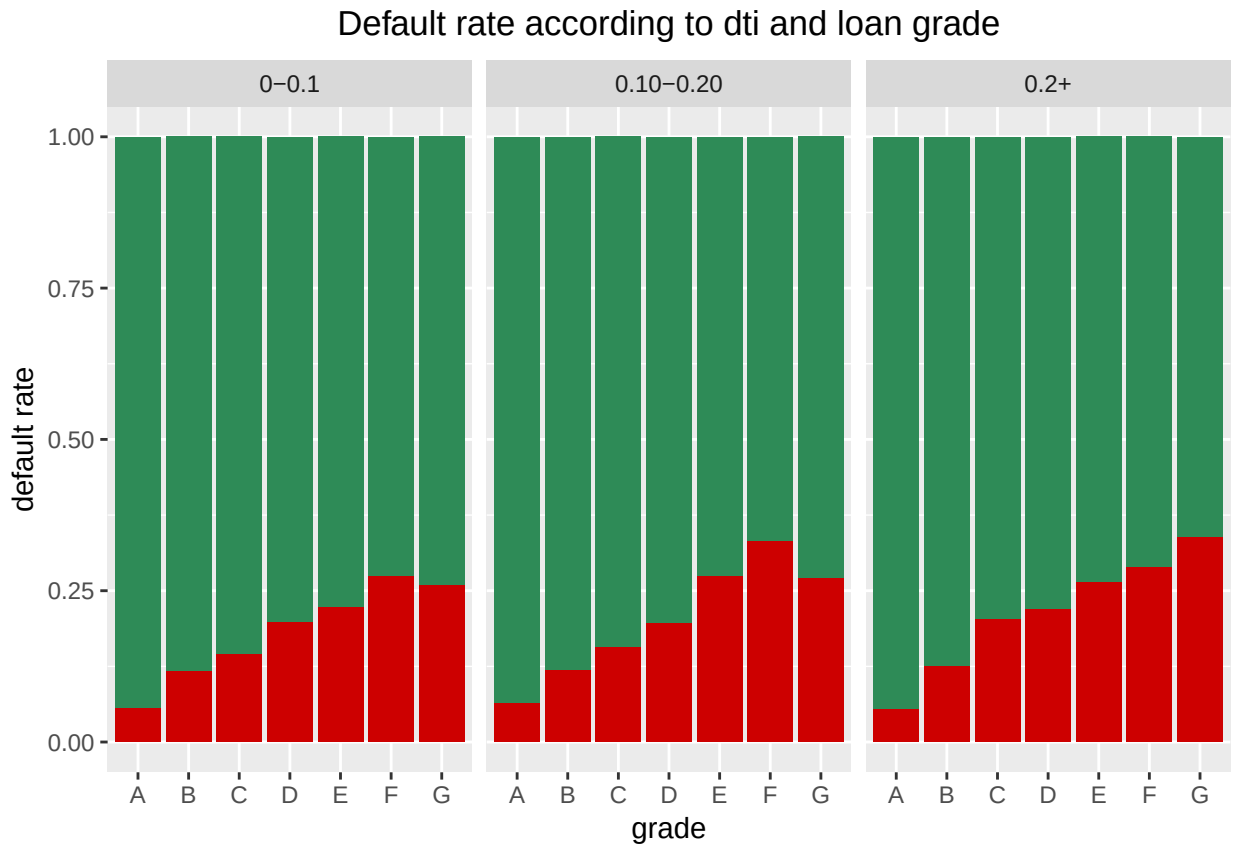
```
model3_0 <- glm(default ~ gradenum + dti, family=binomial, data = Loan0711)
mmps(model3_0)
```




```
binnedplot(fitted(model3_0),residuals(model3_0,type="response"))
```




```
errorrate_model3_0 <- mean((predict(model3_0) > 0.5 & Loan0711$default == 0) |
                           (predict(model3_0) < 0.5 & Loan0711$default == 1))

Loan0711$dtiindex <- ifelse(Loan0711$dti < 0.10,"0-0.1",
                            ifelse(Loan0711$dti <0.20,"0.10-0.20","0.2+"))
ggplot(Loan0711)+aes(x=grade,fill=factor(default))+
geom_bar(position="fill")+facet_grid(~dtiindex)+
labs(title = "Default rate according to dti and loan grade")+
scale_fill_manual(values=c("seagreen4","red3"))+ylab("default rate")+xlab("grade")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")
```



The error rate for the model is 16%.

interpret:

1 sd increase in debt-to-income ratio has a multiple effect of exp(0.01)=1.01 on odds od default,controling grade in same level.

The odds ratio of default for grade A vs grade G is exp(1.8)= 6.05

The loan in grade B is 0.75/4=19% more likely to default than loan in grade A.

The loan in grade C is 1.13/4=28% more likely to default than loan in grade A.

The loan in grade D is 1.38/4=35% more likely to default than loan in grade A.

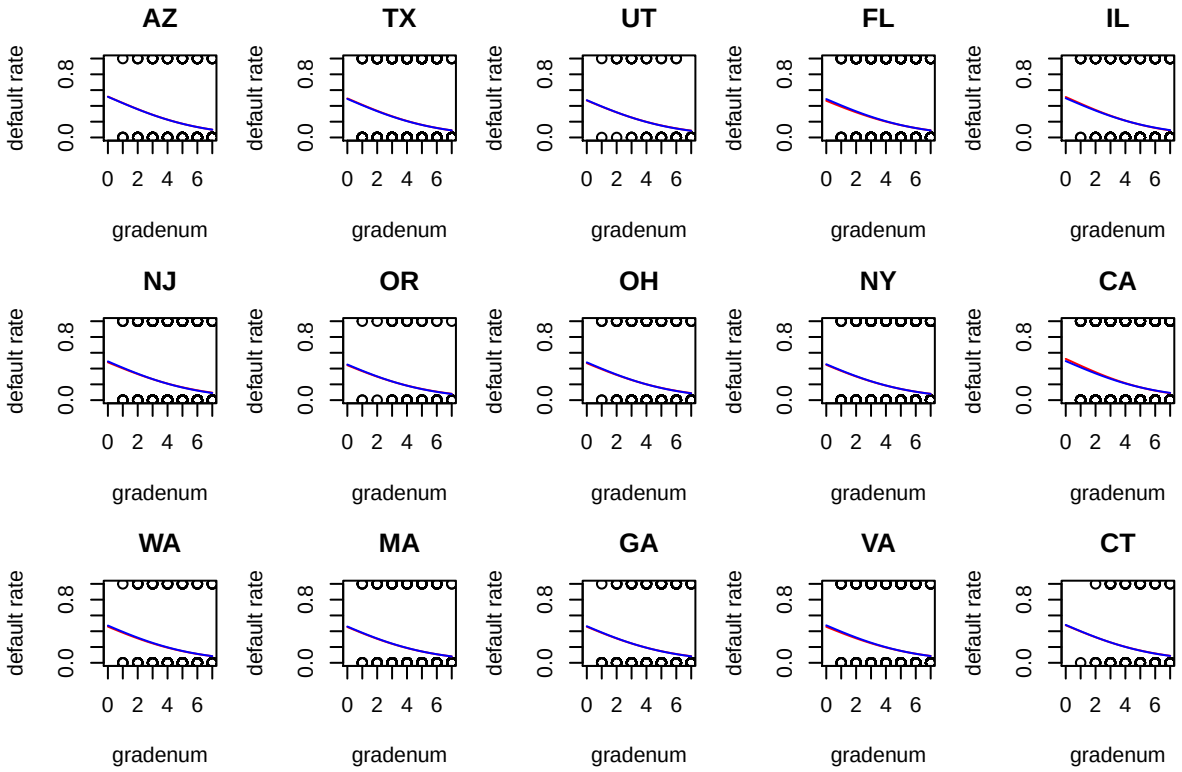
The loan in grade E is 1.70/4=43% more likely to default than loan in grade A.

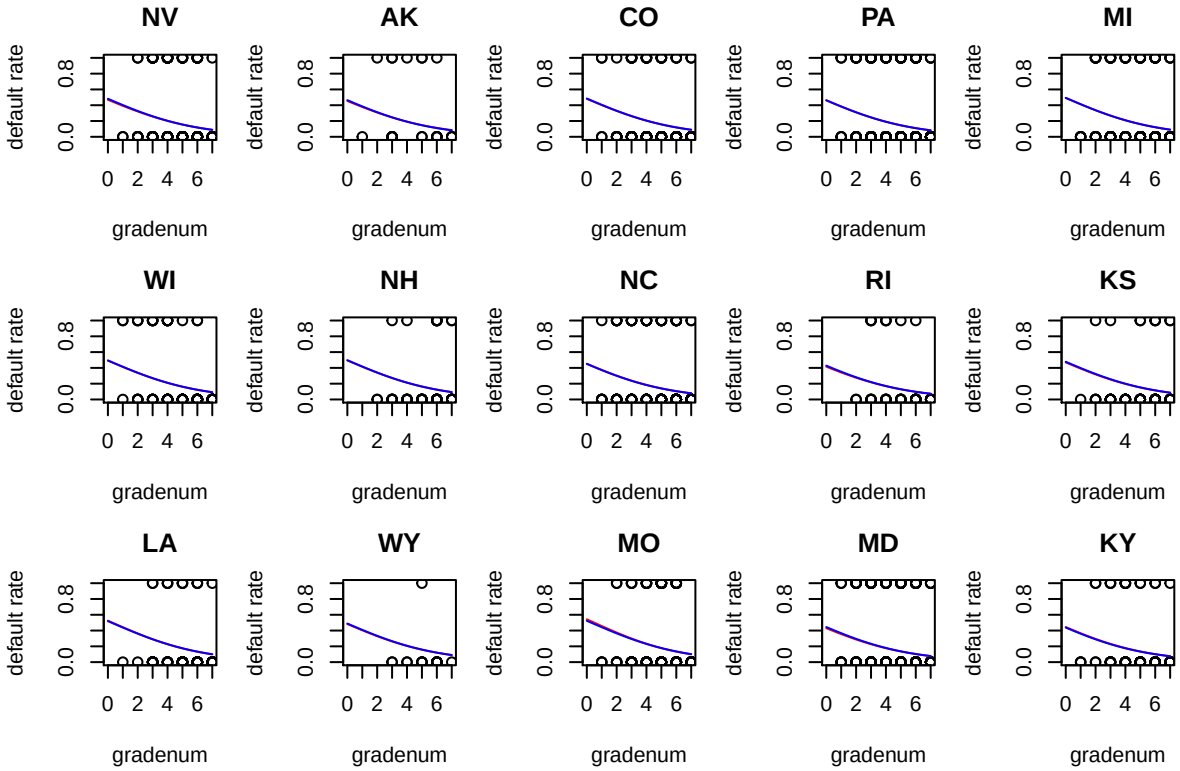
The loan in grade F is 1.94/4=49% more likely to default than loan in grade A.

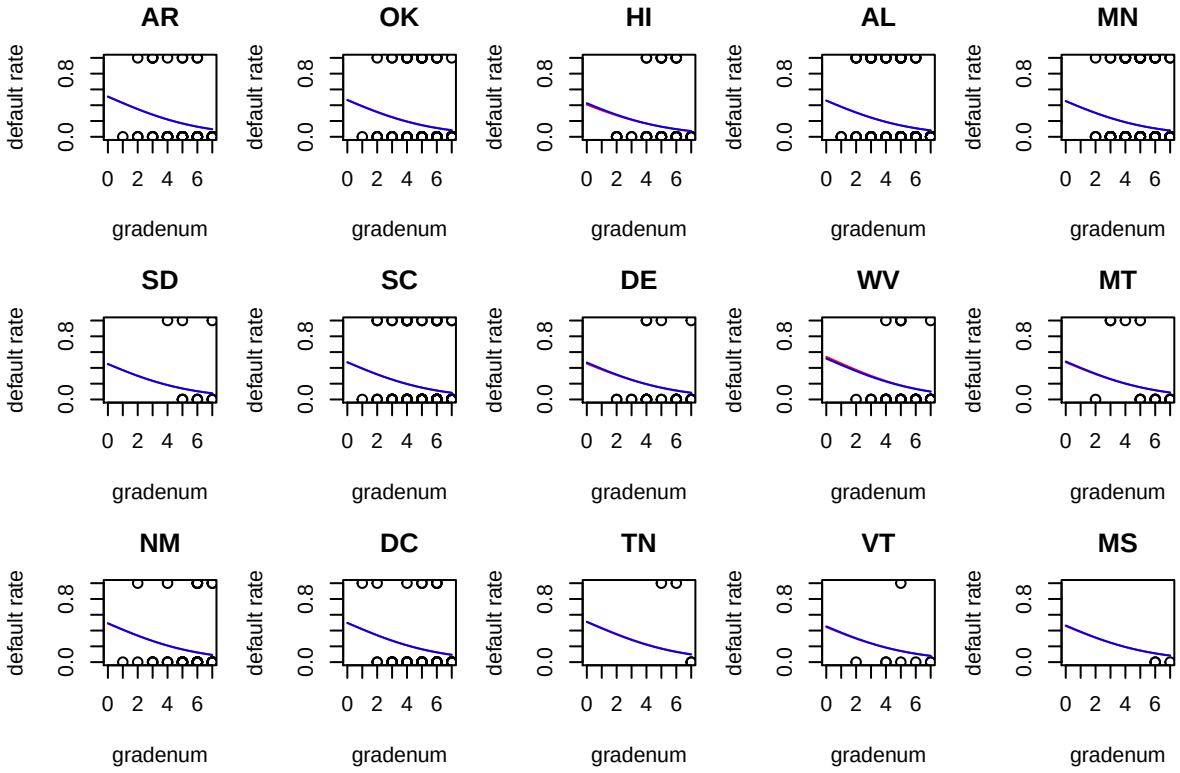
The loan in grade G is 1.82/4=46% more likely to default than loan in grade A.

```
## Multilevel model
# state
model3_2 <- glmer(default ~ gradenum + dti + (1|addr_state),
                  family=binomial(link="logit"), data = Loan0711)
model3_1 <- glmer(default ~ gradenum + dti + (1 + gradenum|addr_state),
                  family=binomial(link="logit"), data = Loan0711)
a.hat.model3_2 <- fixef(model3_2)[1] + ranef(model3_2)$addr_state
b.hat.model3_2 <- fixef(model3_2)[2]
c.hat.model3_2 <- fixef(model3_2)[3]
```

```
a.hat.model3_1 <- fixef(model3_1)[1] + ranef(model3_1)$addr_state[1]
b.hat.model3_1 <- fixef(model3_1)[2] + ranef(model3_1)$addr_state[2]
c.hat.model3_1 <- fixef(model3_1)[3]
x <- Loan0711$gradenum
y <- Loan0711$default
dti <- Loan0711$dti
state <- unique(Loan0711$addr_state)
J <- length(state)
newstate <- rep (NA, J)
for (i in 1:J){newstate[Loan0711$addr_state==state[i]] <- i}
par(mfrow=c(3,5), mar=c(4,4,3,1), oma=c(1,1,2,1))
for (j in 1:J){
plot (x[newstate==j], y[newstate==j],xlim=c(0,7),ylim=c(0,1),
xlab="gradenum", ylab="default rate",main = state[j])
curve (invlogit(a.hat.model3_1[j,] + b.hat.model3_1[j,]*x + c.hat.model3_1*dti[j]),
       col="red", add=TRUE)
curve (invlogit(a.hat.model3_2[j,] + b.hat.model3_2*x + c.hat.model3_2*dti[j]),
       col="blue", add=TRUE)
}
```

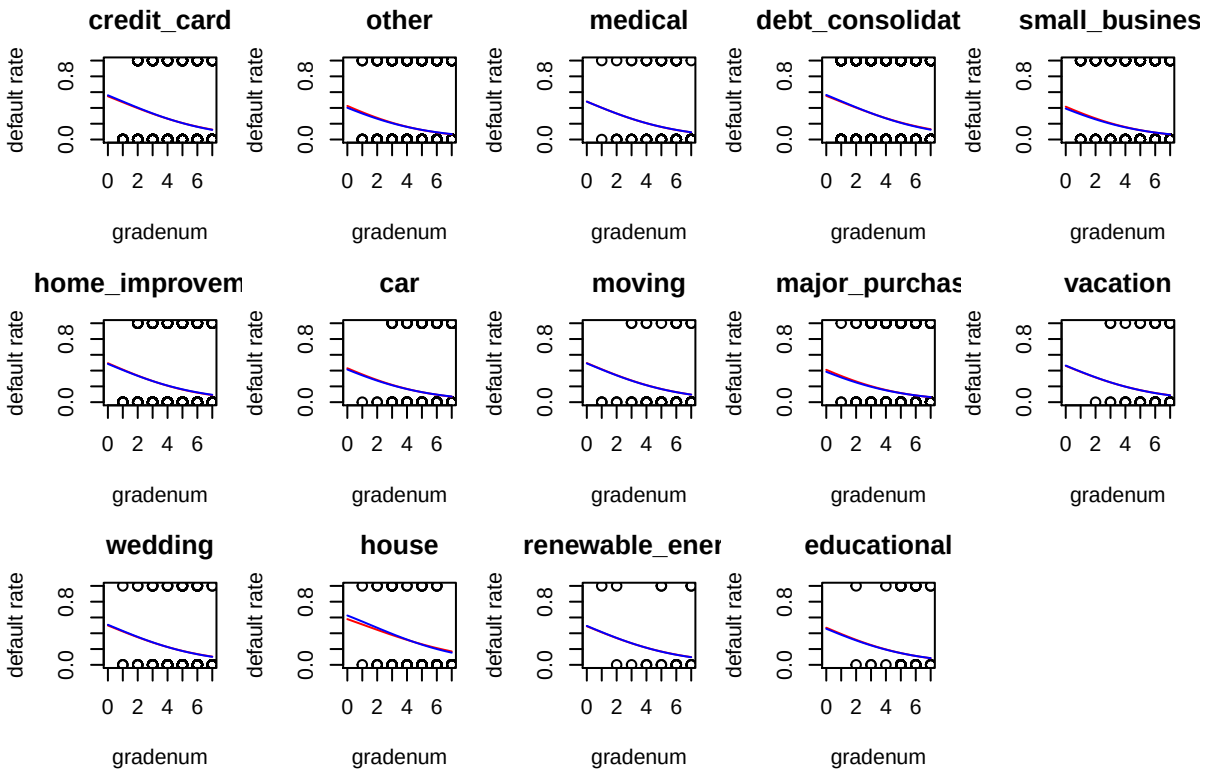



NV
AK
CO
PA
MI
WI
NH
NC
RI
KS
LA
WY
MO
MD
KY
default rate
0.0
0.8
0 2 4 6
gradenum




```
# purpose
model4_2 <- glmer(default ~ gradenum + dti + (1|purpose),
                  family=binomial(link="logit"), data = Loan0711)
model4_1 <- glmer(default ~ gradenum + dti + (1 + gradenum|purpose),
                  family=binomial(link="logit"), data = Loan0711)
a.hat.model4_2 <- fixef(model4_2)[1] + ranef(model4_2)$purpose
b.hat.model4_2 <- fixef(model4_2)[2]
c.hat.model4_2 <- fixef(model4_2)[3]

a.hat.model4_1 <- fixef(model4_1)[1] + ranef(model4_1)$purpose[1]
b.hat.model4_1 <- fixef(model4_1)[2] + ranef(model4_1)$purpose[2]
c.hat.model4_1 <- fixef(model4_1)[3]
x <- Loan0711$gradenum
y <- Loan0711$default
dti <- Loan0711$dti
purpose <- unique(Loan0711$purpose)
J <- length(purpose)
newpurpose <- rep (NA, J)
for (i in 1:J){newpurpose[Loan0711$purpose==purpose[i]] <- i}
par(mfrow=c(3,5), mar=c(4,4,3,1), oma=c(1,1,2,1))
for (j in 1:J){
plot (x[newpurpose==j], y[newpurpose==j],xlim=c(0,7),ylim=c(0,1),
xlab="gradenum", ylab="default rate",main = purpose[j])
curve (invlogit(a.hat.model4_1[j,] + b.hat.model4_1[j,]*x + c.hat.model4_1*dti[j]),
       col="red", add=TRUE)
curve (invlogit(a.hat.model4_2[j,] + b.hat.model4_2*x + c.hat.model4_2*dti[j]),
```

```
        col="blue", add=TRUE)
}
```



4. Discussion

1. Result

From this analysis report, we could know that the loan grade is significantly related to loan amount and loan term. What's more, if borrower had another mortgage, the grade would be higher as expected. This is reasonable, since their credit quality had checked by other institutions. However, the employment year is not such related as common sense, and this is a point that could be analyze deeper.

For default rate, the higher grade reflect the lower default rate. What's more, the initial debt to income ratio is also important. From the analysis, we could give suggestion that grade F and grade G has really high default rate, and reject their loan request may be a better desicion.

2. Data limitation

Since the credit information is private, we cannot track the whole process for each loan.

What's more, the FICO Score should be an important predictor in this analysis,however, since it is personal privacy, we could not get the data. Therefore, just consider about the factor as home ownership and income has its limitation.

Finally, since the financial market and the interest rate policy has changed during years, credit quality could not be the only factor of default. Other investment chances and bank loan should be considered and the time span is important.

3. Future directions

Combine the bank interest rate in each years analysis is what we could consider deeper, and basic financial knowledge is required. Moreover, since the loan is always span 3-5 years or more, maybe we could track the result years later.

Appendix and reference

https://www.lendingclub.com/info/download-data.action