loan\_approval\_analysis

ANKIT

2025-09-07

# TITLE= LOAN APPROVAL ANALYSIS  
#AUTHOR- ANKIT GHOSH  
  
  
#import libraries   
library(tidyr)  
library(dplyr)  
library(ggplot2)  
  
  
#data import  
setwd("D:/r\_project/r\_islr\_work/LOAN\_APPROVAL\_ANALYSIS")  
getwd()

## [1] "D:/r\_project/r\_islr\_work/LOAN\_APPROVAL\_ANALYSIS"

df=read.csv("Loan\_Train.csv")  
dim(df)

## [1] 614 13

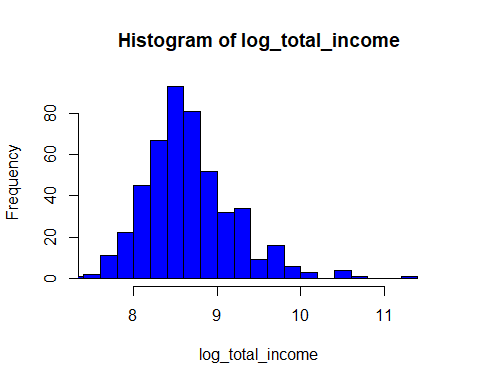
#data cleaning  
df[df==""]=NA  
df1=drop\_na(df)  
dim(df1)

## [1] 480 13

str(df1)

## 'data.frame': 480 obs. of 13 variables:  
## $ Loan\_ID : chr "LP001003" "LP001005" "LP001006" "LP001008" ...  
## $ Gender : chr "Male" "Male" "Male" "Male" ...  
## $ Married : chr "Yes" "Yes" "Yes" "No" ...  
## $ Dependents : chr "1" "0" "0" "0" ...  
## $ Education : chr "Graduate" "Graduate" "Not Graduate" "Graduate" ...  
## $ Self\_Employed : chr "No" "Yes" "No" "No" ...  
## $ ApplicantIncome : int 4583 3000 2583 6000 5417 2333 3036 4006 12841 3200 ...  
## $ CoapplicantIncome: num 1508 0 2358 0 4196 ...  
## $ LoanAmount : int 128 66 120 141 267 95 158 168 349 70 ...  
## $ Loan\_Amount\_Term : int 360 360 360 360 360 360 360 360 360 360 ...  
## $ Credit\_History : int 1 1 1 1 1 1 0 1 1 1 ...  
## $ Property\_Area : chr "Rural" "Urban" "Urban" "Urban" ...  
## $ Loan\_Status : chr "N" "Y" "Y" "Y" ...

Total\_Income=df1$ApplicantIncome+df1$CoapplicantIncome  
log\_total\_income=log(Total\_Income,base = exp(1))  
hist(log\_total\_income,col="blue",xlim = c(7.5,11.5),breaks = 20)



#removing oulier for applicant income  
quantile(df1$ApplicantIncome)

## 0% 25% 50% 75% 100%   
## 150.00 2898.75 3859.00 5852.50 81000.00

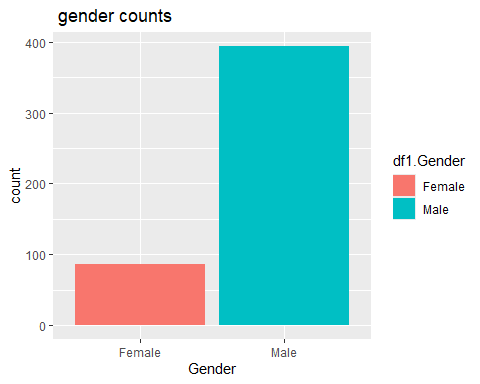
u=IQR(df1$ApplicantIncome)  
u

## [1] 2953.75

u\_b=5852.50+1.5\*u  
l\_b=2898.75 -1.5\*u  
df2=subset(df1,df1$ApplicantIncome>l\_b & df1$ApplicantIncome<u\_b)  
  
  
#PERFORMING SOME SORTS OF EDA RELATED TO DATA  
  
#number of people who take loan group by gender  
loan\_gender=data.frame(df1$Gender)  
gender\_count=loan\_gender%>%count(df1$Gender)  
print(gender\_count)

## df1$Gender n  
## 1 Female 86  
## 2 Male 394

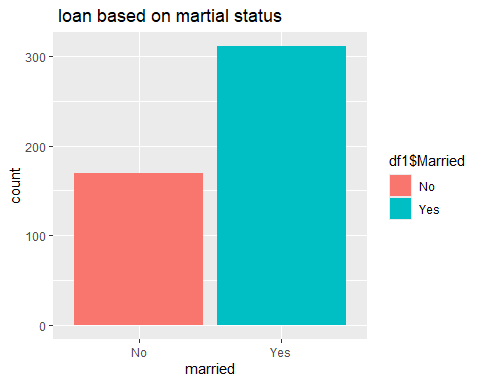
ggplot(data = loan\_gender,aes(x=df1$Gender))+geom\_bar(aes(fill = df1.Gender))+labs(title=" gender counts",x="Gender")+theme()



#number of people who take loan group by martial status  
loan\_married=data.frame(df1$Married)  
married\_count=loan\_married%>%count(df1$Married)  
print(married\_count)

## df1$Married n  
## 1 No 169  
## 2 Yes 311

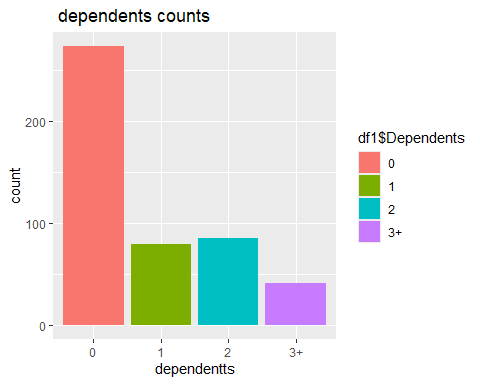
ggplot(data = loan\_married,aes(x=df1.Married))+geom\_bar(aes(fill=df1$Married))+labs(title=" loan based on martial status",x="married")+theme()



#number of people who take loan group by dependents  
loan\_dependents=data.frame(df1$Dependents)  
dependents\_count=loan\_dependents%>%count(df1$Dependents)  
print(dependents\_count)

## df1$Dependents n  
## 1 0 274  
## 2 1 80  
## 3 2 85  
## 4 3+ 41

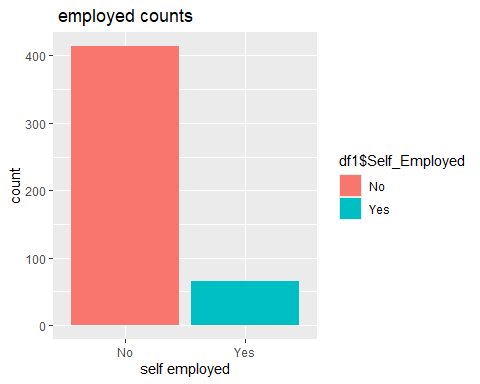
ggplot(data = loan\_dependents,aes(x=df1.Dependents))+geom\_bar(aes(fill=df1$Dependents))+labs(title=" dependents counts",x="dependentts")+theme()



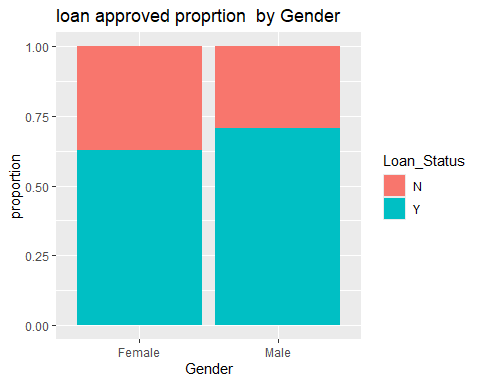
#number of people who take loan group by self employed  
loan\_employed=data.frame(df1$Self\_Employed)  
employed\_count=loan\_employed%>%count(df1$Self\_Employed)  
print(employed\_count)

## df1$Self\_Employed n  
## 1 No 414  
## 2 Yes 66

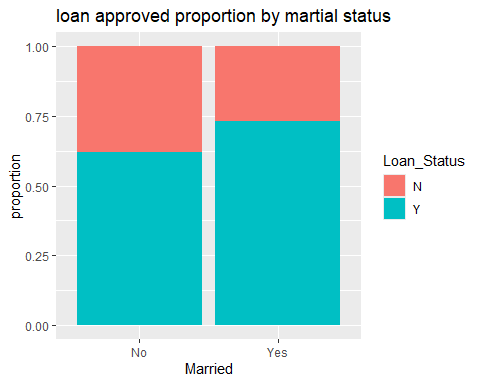
ggplot(data = loan\_employed,aes(x=df1.Self\_Employed))+geom\_bar(aes(fill=df1$Self\_Employed))+labs(title=" employed counts",x="self employed")+theme()



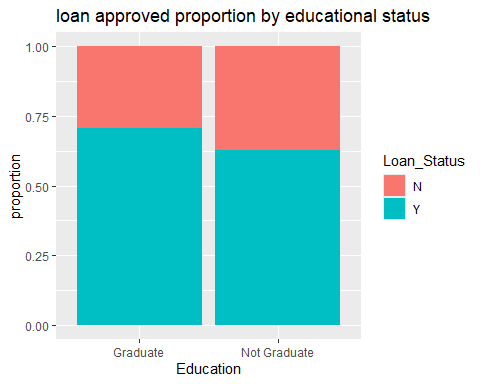
#proportion of loan approved by demographic analysis  
  
#loan approved proportion by gender  
ggplot(df1,aes(x=Gender,fill=Loan\_Status))+geom\_bar(position = "fill")+labs(title = "loan approved proprtion by Gender",y="proportion")



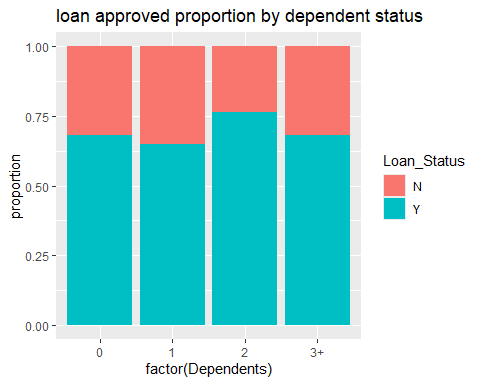
#loan approved proportion by martial status  
ggplot(df1,aes(x=Married,fill=Loan\_Status))+geom\_bar(position = "fill")+labs(title = "loan approved proportion by martial status",y="proportion")



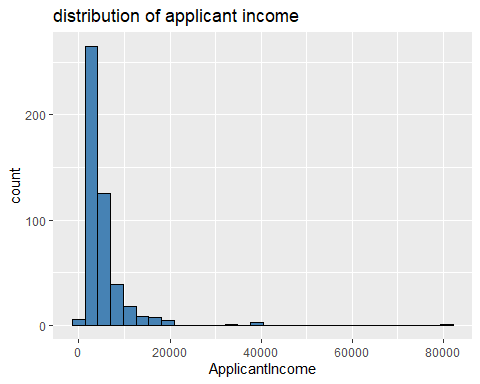
#loan approved proportion by educational status  
ggplot(df1,aes(x=Education,fill=Loan\_Status))+geom\_bar(position = "fill")+labs(title = "loan approved proportion by educational status",y="proportion")



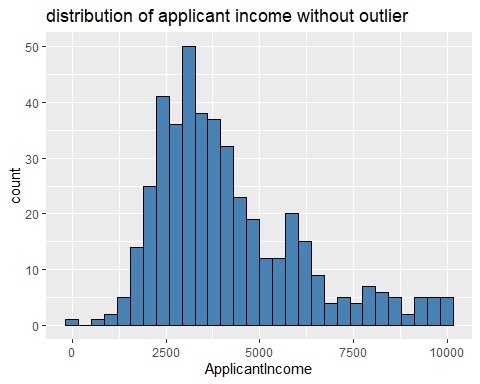
#loan approved proportion by dependent status  
ggplot(df1,aes(x=factor(Dependents),fill=Loan\_Status))+geom\_bar(position = "fill")+labs(title = "loan approved proportion by dependent status",y="proportion")



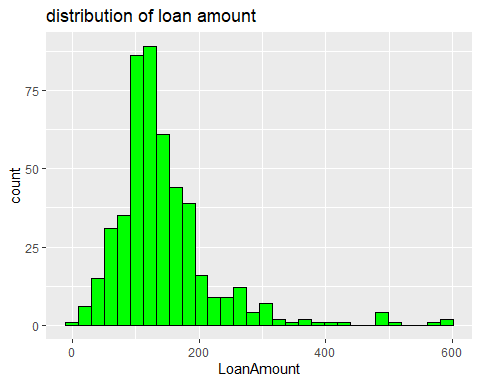
#income and loan analysis  
  
#income distribution  
ggplot(df1,aes(x=ApplicantIncome))+geom\_histogram(bins = 30,fill="steelblue",color="black")+labs(title = "distribution of applicant income")



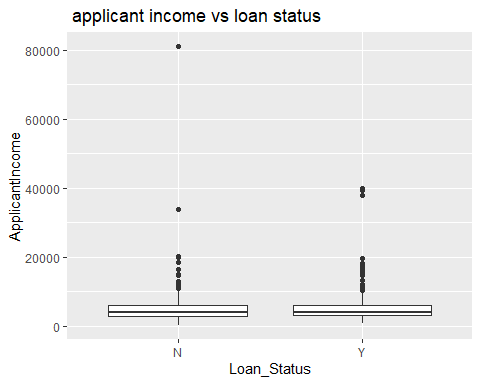
#income distribution after removing outlier  
ggplot(df2,aes(x=ApplicantIncome))+geom\_histogram(bins = 30,fill="steelblue",color="black")+labs(title = "distribution of applicant income without outlier")



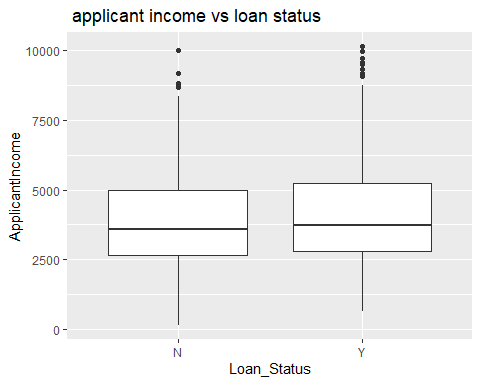
#loan amount distribution  
ggplot(df1,aes(x=LoanAmount))+geom\_histogram(bins = 30,fill="green",color="black")+labs(title = "distribution of loan amount")



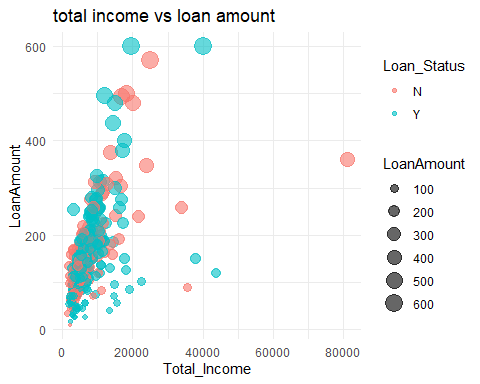
#income vs loanstatus  
ggplot(df1,aes(x=Loan\_Status,y=ApplicantIncome))+geom\_boxplot()+labs(title = " applicant income vs loan status")



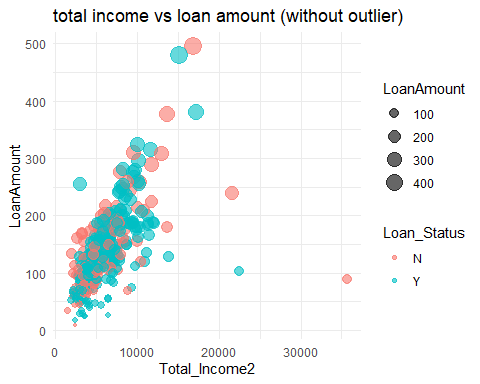
#income vs loanstatus after removing outlier  
ggplot(df2,aes(x=Loan\_Status,y=ApplicantIncome))+geom\_boxplot()+labs(title = " applicant income vs loan status")



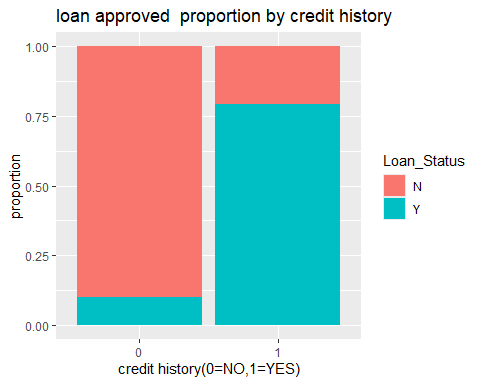
#total income vs loan amount  
ggplot(df1,aes(x=Total\_Income,y=LoanAmount,color=Loan\_Status))+geom\_point(aes(size=LoanAmount),alpha=0.6)+labs(title = "total income vs loan amount")+theme\_minimal()



#total income vs loan amount after removing outlier  
Total\_Income2=df2$ApplicantIncome+df2$CoapplicantIncome  
ggplot(df2,aes(x=Total\_Income2,y=LoanAmount,color=Loan\_Status))+geom\_point(aes(size=LoanAmount),alpha=0.6)+labs(title = "total income vs loan amount (without outlier)")+theme\_minimal()



#CREDIT HISTORY AND LOAN APPROVAL  
ggplot(df1,aes(x=factor(Credit\_History),fill = Loan\_Status))+geom\_bar(position = "fill")+labs(title = "loan approved proportion by credit history",x="credit history(0=NO,1=YES)",y="proportion")



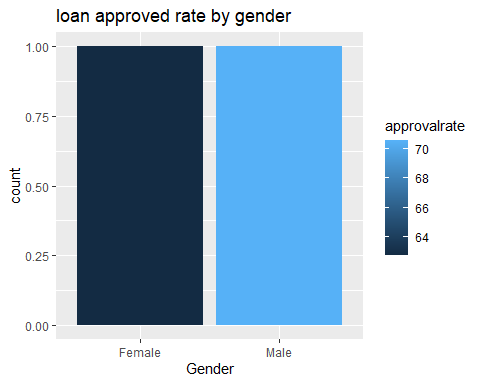
#LOAN APPROVAL RATE OVERALL  
approv\_rate =mean(df1$Loan\_Status=="Y")\*100  
cat("overall loan approval rate:",round(approv\_rate,2),"%\n")

## overall loan approval rate: 69.17 %

#Approval rate by gender  
gen\_approv=df1%>%group\_by(Gender)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(gen\_approv)

## # A tibble: 2 × 2  
## Gender approvalrate  
## <chr> <dbl>  
## 1 Female 62.8  
## 2 Male 70.6

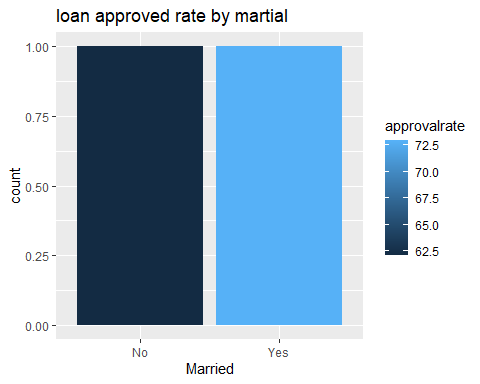
ggplot(gen\_approv,aes(x=Gender,fill =approvalrate ))+geom\_bar(position = "fill")+labs(title = "loan approved rate by gender")



#approval rate by martial status   
married\_approval=df1%>%group\_by(Married)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(married\_approval)

## # A tibble: 2 × 2  
## Married approvalrate  
## <chr> <dbl>  
## 1 No 62.1  
## 2 Yes 73.0

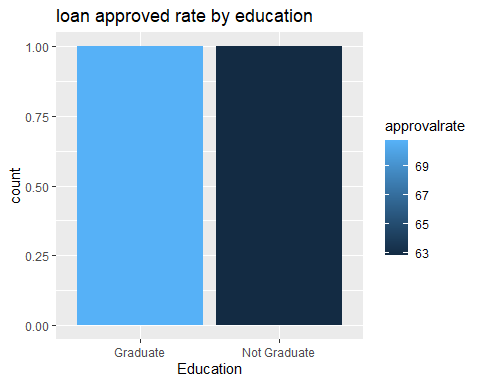
ggplot(married\_approval,aes(x=Married,fill =approvalrate ))+geom\_bar(position = "fill")+labs(title = "loan approved rate by martial ")



#approval rate by educational status   
edu\_approval=df1%>%group\_by(Education)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(edu\_approval)

## # A tibble: 2 × 2  
## Education approvalrate  
## <chr> <dbl>  
## 1 Graduate 70.8  
## 2 Not Graduate 62.9

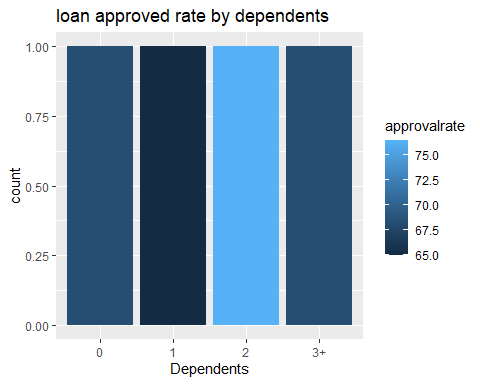
ggplot(edu\_approval,aes(x=Education,fill =approvalrate ))+geom\_bar(position = "fill")+labs(title = "loan approved rate by education")



#approval rate by Dependents status   
dep\_approval=df1%>%group\_by(Dependents)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(dep\_approval)

## # A tibble: 4 × 2  
## Dependents approvalrate  
## <chr> <dbl>  
## 1 0 68.2  
## 2 1 65   
## 3 2 76.5  
## 4 3+ 68.3

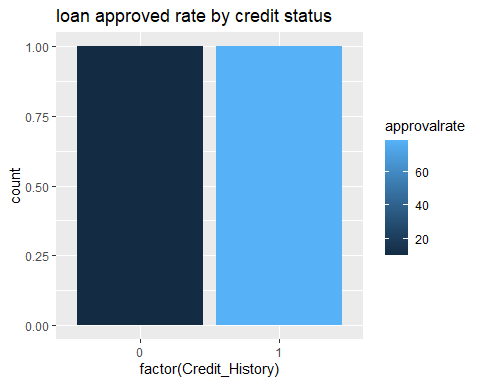
ggplot(dep\_approval,aes(x=Dependents,fill =approvalrate ))+geom\_bar(position = "fill")+labs(title = "loan approved rate by dependents")



#approval rate by credit status   
cred\_approval=df1%>%group\_by(Credit\_History)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(cred\_approval)

## # A tibble: 2 × 2  
## Credit\_History approvalrate  
## <int> <dbl>  
## 1 0 10   
## 2 1 79.3

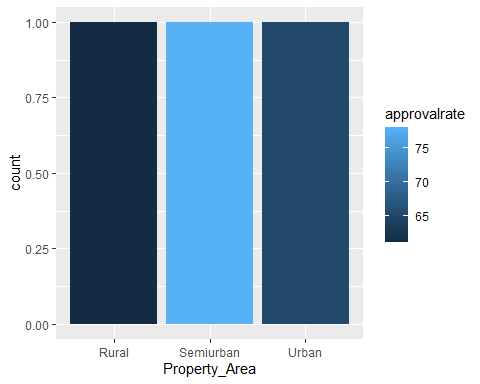
ggplot(cred\_approval,aes(x=factor(Credit\_History),fill = approvalrate ))+geom\_bar(position = "fill")+labs(title = "loan approved rate by credit status")



#approval rate by property area   
p\_approval=df1%>%group\_by(Property\_Area)%>%summarise(approvalrate=mean(Loan\_Status=="Y",na.rm=TRUE)\*100)  
print(p\_approval)

## # A tibble: 3 × 2  
## Property\_Area approvalrate  
## <chr> <dbl>  
## 1 Rural 61.2  
## 2 Semiurban 78.0  
## 3 Urban 65.3

ggplot(p\_approval,aes(x=Property\_Area,fill =approvalrate ))+geom\_bar(position = "fill")



#average loan amount by loan status  
avg\_loan=df1%>%group\_by(Loan\_Status)%>%summarise(AverageLoan=mean(LoanAmount,na.rm=TRUE))  
print(avg\_loan)

## # A tibble: 2 × 2  
## Loan\_Status AverageLoan  
## <chr> <dbl>  
## 1 N 153.  
## 2 Y 141.

ggplot(avg\_loan,aes(x=Loan\_Status,fill =AverageLoan ))+geom\_bar(position = "fill")+labs(title = "average loan vs loan status")

