EDA01.R

chidam

Sat Aug 29 17:07:54 2015

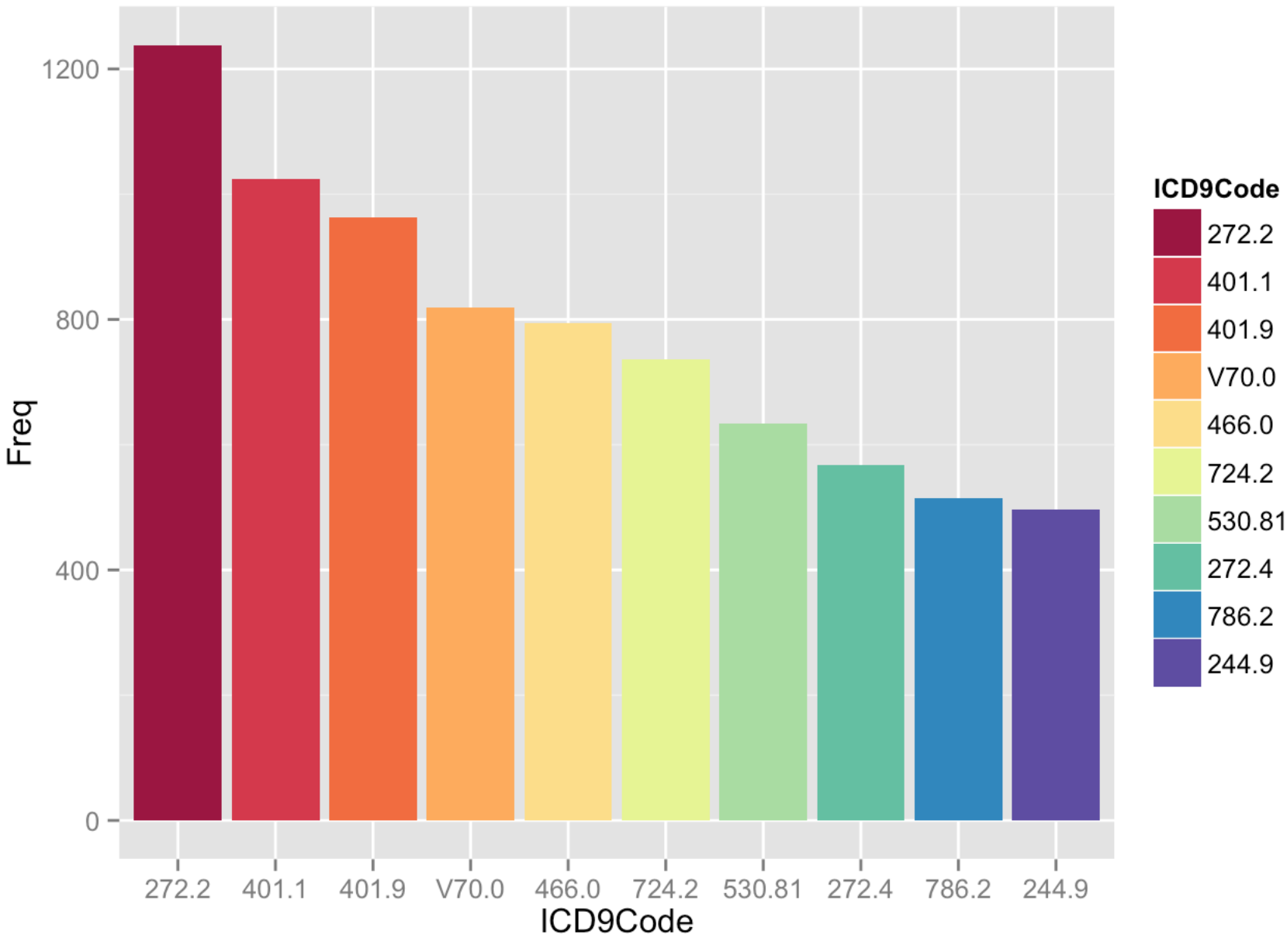
```
sample_diag=Diagnosis
sample_diag$ICD9Code=as.factor(sample_diag$ICD9Code)
head(sort(table(sample_diag$ICD9Code),decreasing=TRUE),10)
```

```
##
##  272.2  401.1  401.9  V70.0  466.0  724.2 530.81  272.4  786.2  244.9
##   1238   1024    963    819    795    736    634    568    515    496
```

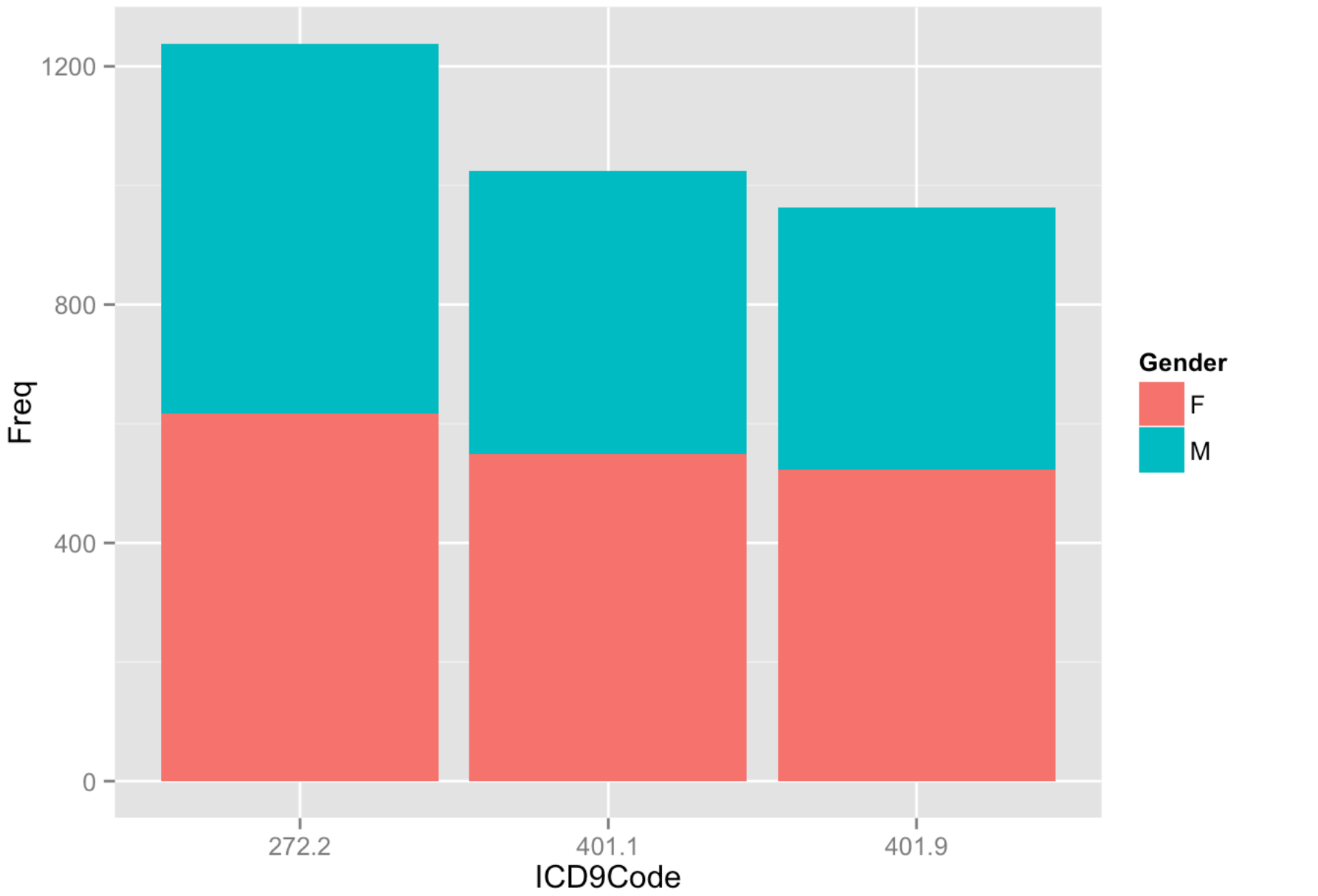
```
top10diseasesCounts=as.matrix(head(sort(table(sample_diag$ICD9Code),decreasing=TRUE),
10))[1:10]
top10diseasesCodes=dimnames(as.matrix(head(sort(table(sample_diag$ICD9Code),decreasin
g=TRUE),10)))[[1]]
top10diseasesDF=data.frame(top10diseasesCodes,top10diseasesCounts)
top10diseasesDF
```

```
##    top10diseasesCodes top10diseasesCounts
## 1               272.2                1238
## 2               401.1                1024
## 3               401.9                 963
## 4               V70.0                 819
## 5               466.0                 795
## 6               724.2                 736
## 7              530.81                 634
## 8               272.4                 568
## 9               786.2                 515
## 10              244.9                 496
```

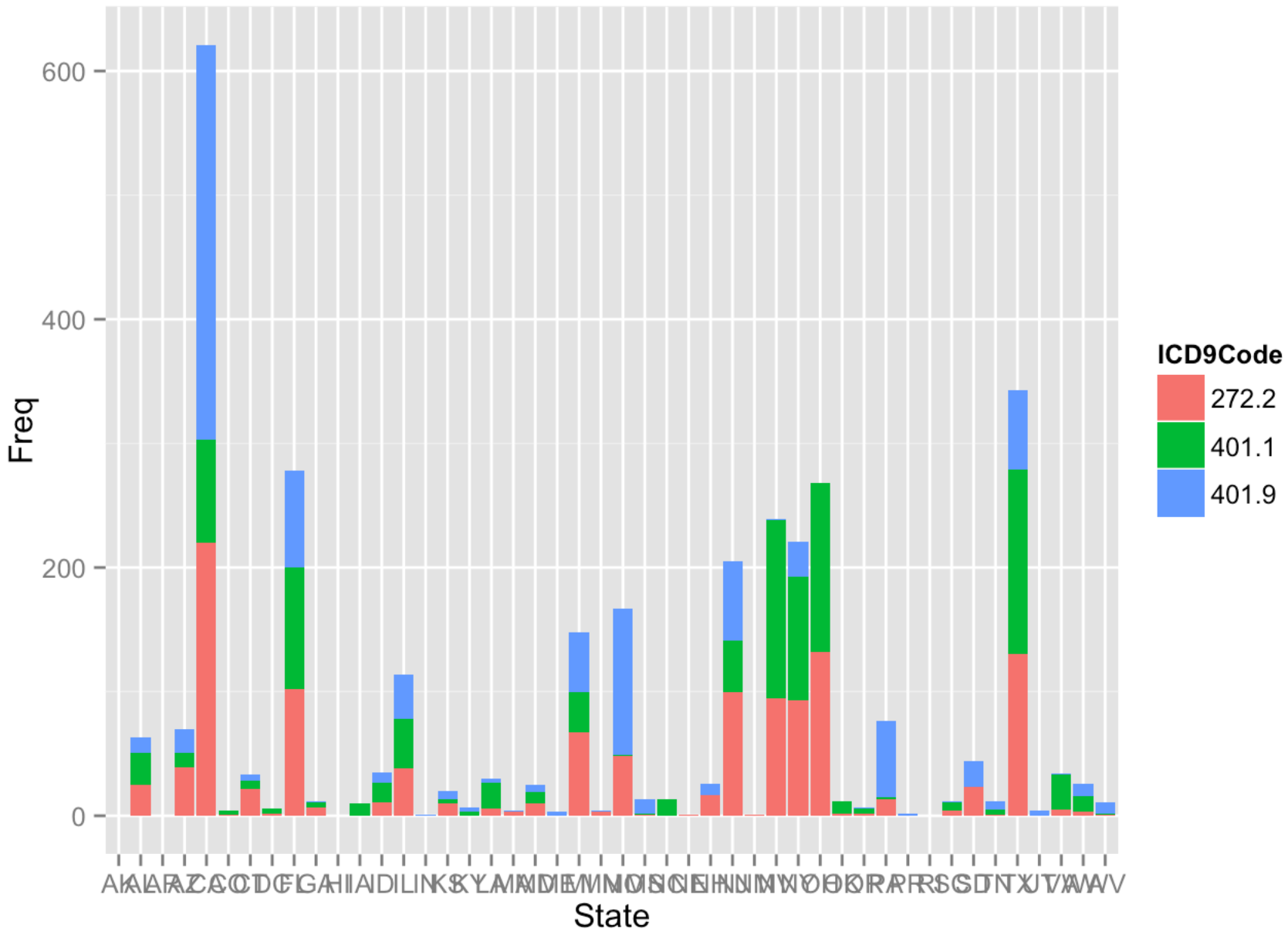
```
names(top10diseasesDF)=c('ICD9Code','Freq')
res=mapply(function(d){
  head(which(sample_diag$ICD9Code==d),1)
},as.character(top10diseasesDF$ICD9Code))
res_diag=mapply(function(dd){
  sample_diag$DiagnosisDescription[dd]
  },res)
rank1=1:10
disease_des=as.matrix(res_diag)[1:10]
top10diseasesDF<-cbind(top10diseasesDF,disease_des,rank1)
top10diseasesDF$ICD9Code=factor(top10diseasesDF$ICD9Code,levels=top10diseasesDF$ICD9C
ode[order(top10diseasesDF$rank1)])
set.seed(10)
library(ggplot2)
##Build Pareto chart of top 10 diseases
ggplot(top10diseasesDF,aes(x=ICD9Code,y=Freq,fill=ICD9Code))+geom_bar(stat = "identit
y")+
#scale_colour_gradientn(colours=rainbow(4))
  scale_fill_brewer(palette="Spectral")
```



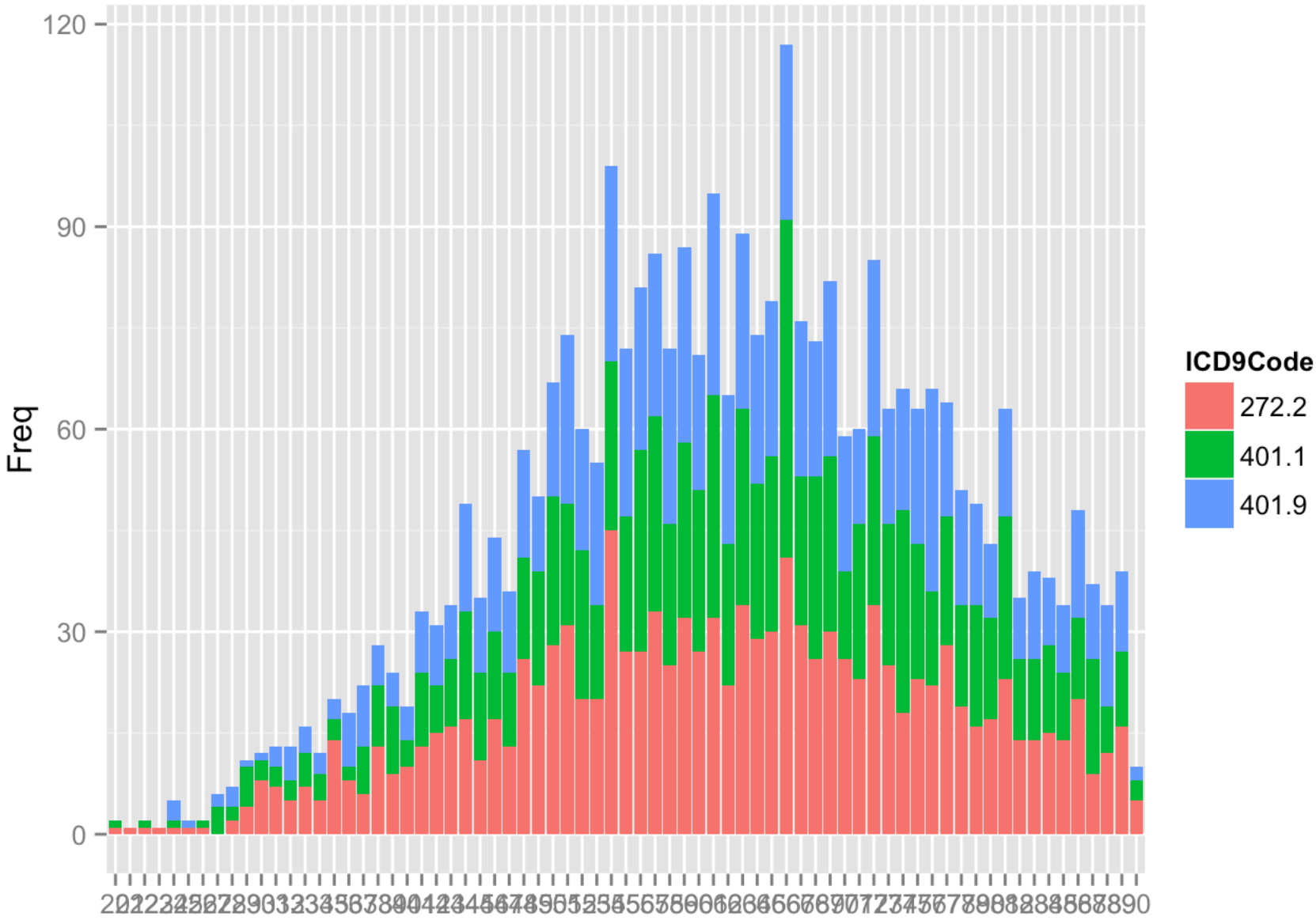
```
#selecting top 3 diseases
top3diseasesDF=head(top10diseasesDF,3)
#Gender, State & Age wise split
Patient_Diagnosis1=Patient_Diagnosis
Patient_Diagnosis1$age=2012-as.numeric(Patient_Diagnosis1$YearOfBirth)
Patient_Diagnosis2=subset(Patient_Diagnosis1,ICD9Code %in% top3diseasesDF$ICD9Code)
Patients_with_hypTension=unique(Patient_Diagnosis2$PatientGuid)

#Studying diseases based on Gender
Patient_Diagnosis3=as.data.frame(as.matrix(table(Patient_Diagnosis2$Gender,Patient_Di
agnosis2$ICD9Code)))
names(Patient_Diagnosis3)=c('Gender','ICD9Code','Freq')
ggplot(Patient_Diagnosis3,aes(x=ICD9Code,y=Freq,fill=Gender,label="Top 3 ICD9 disease
based on gender"))+geom_bar(stat = "identity")
```



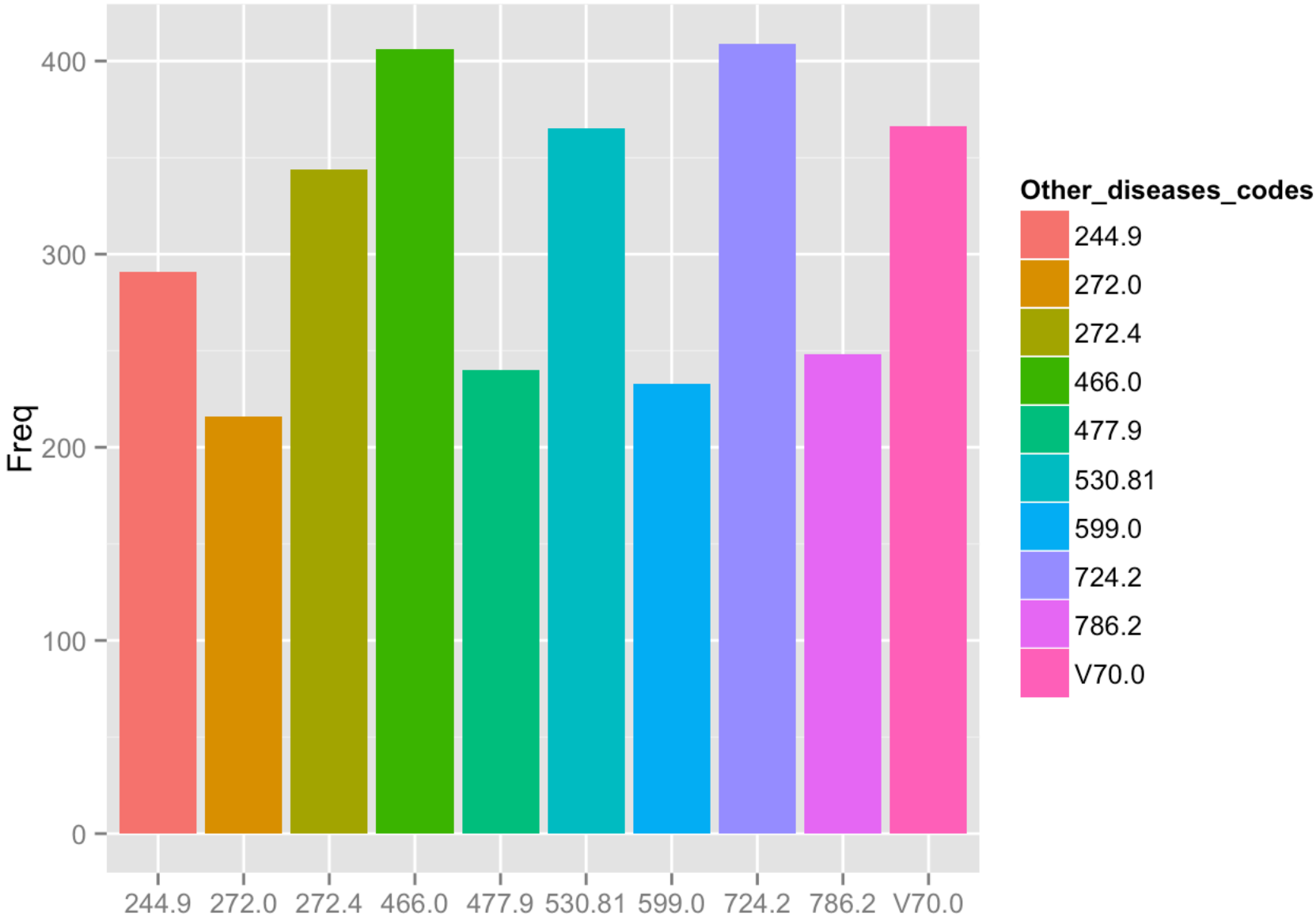

```
#Studying diseases across states in US
Patient_Diagnosis4=as.data.frame(as.matrix(table(Patient_Diagnosis2$State,Patient_Dia
gnosis2$ICD9Code)))
names(Patient_Diagnosis4)=c('State','ICD9Code','Freq')
ggplot(Patient_Diagnosis4,aes(x=State,y=Freq,fill=ICD9Code,label="Top 3 ICD9 disease
based across States"))+geom_bar(stat = "identity")
```




```
#Studying diseases based on Age of the patients
Patient_Diagnosis5=as.data.frame(as.matrix(table(Patient_Diagnosis2$age,Patient_Diagn
osis2$ICD9Code)))
names(Patient_Diagnosis5)=c('age','ICD9Code','Freq')
ggplot(Patient_Diagnosis5,aes(x=age,y=Freq,fill=ICD9Code,label="Top 3 ICD9 disease ba
sed across States"))+geom_bar(stat = "identity")
```




```
#How Hypertension translates into other Disease?
Patient_Diagnosis_other=subset(Patient_Diagnosis1,!(ICD9Code %in% top3diseasesDF$ICD9
Code))
Patient_with_hypTension_and_others=subset(Patient_Diagnosis_other,PatientGuid %in% Pa
tients_with_hypTension)
Other_diseases_count=as.data.frame(as.matrix(head(sort(table(Patient_with_hypTension_
and_others$ICD9Code),decreasing=TRUE),10)))[[1]]
Other_diseases_codes=row.names(as.data.frame(as.matrix(head(sort(table(Patient_with_h
ypTension_and_others$ICD9Code),decreasing=TRUE),10))))
Other_diseases=data.frame(Other_diseases_codes,Other_diseases_count)
names(Other_diseases)=c('Other_diseases_codes','Freq')
ggplot(Other_diseases,aes(x=Other_diseases_codes,y=Freq,fill=Other_diseases_codes,lab
el="HyperTension patients' other diseases"))+geom_bar(stat = "identity")
```


Freq
400
300
200
100
0
244.9
272.0
272.4
466.0
477.9
530.81
599.0
724.2
786.2
V70.0
Other_diseases_codes
Other_diseases_codes
244.9
272.0
272.4
466.0
477.9
530.81
599.0
724.2
786.2
V70.0