Challenge Solution

Part I

challenge 1

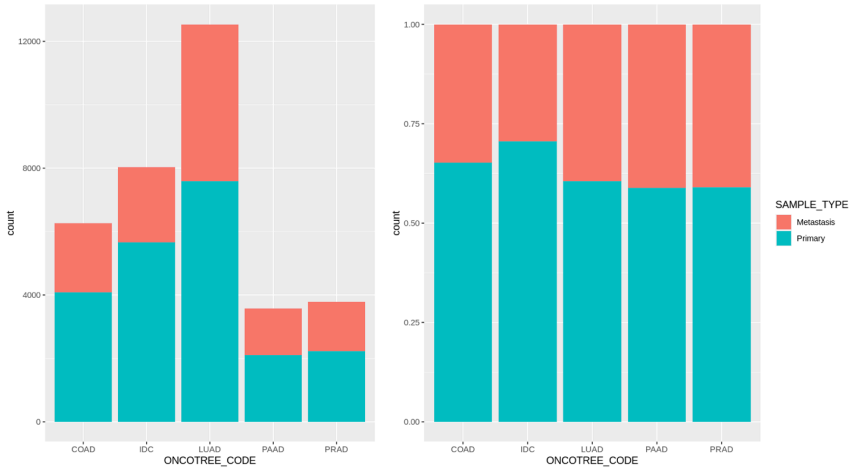
```
## in  R

#merge sample and patient information
sample <- as.data.frame(read.table("data_clinical_sample.txt",header = TRUE,sep
= "\t", dec = ".",stringsAsFactors=FALSE,check.names = FALSE))
patients <- as.data.frame(read.table("data_clinical_patient.txt",header =
TRUE,sep = "\t", dec = ".",stringsAsFactors=FALSE,check.names = FALSE))
dat<-merge(sample,patients[,c("PATIENT_ID","SEX")],by="PATIENT_ID",all.x=TRUE)
dim(dat)

#select target cancer type tumor samples
dat=subset(dat,SAMPLE_TYPE=="Primary"|SAMPLE_TYPE=="Metastasis")
dat=subset(dat,ONCOTREE_CODE=="LUAD"|ONCOTREE_CODE=="IDC"|ONCOTREE_CODE=="COAD"
|ONCOTREE_CODE=="PRAD"|ONCOTREE_CODE=="PAAD")
#dat=subset(dat,ONCOTREE_CODE=="LUAD")
head(dat)
dim(dat)

#calculate sample type frequence in each cancer type
library(dplyr)
summaryData<-as.data.frame(group_by(dat,ONCOTREE_CODE,SAMPLE_TYPE) %>%
summarise(.,count=n()))
summaryData

#plot
par(mfrow=c(1,2))
ggplot(summaryData,aes(x=ONCOTREE_CODE,
y=count,fill=SAMPLE_TYPE))+geom_bar(stat='identity')
ggplot(summaryData,aes(x=ONCOTREE_CODE, y=count,
fill=SAMPLE_TYPE))+geom_bar(stat='identity',position="fill")
```

112935 · 10

A data.frame: 6 x 10

	PATIENT_ID	SAMPLE_ID	AGE_AT_SEQ_REPORT	ONCOTREE_CODE	SAMPLE_TYPE	SEQ_ASSAY_ID	CANCER_TYPE	CANCER_TYPE_DETAILED	SAMPLI
	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	
835	GENIE-COLU-00206	GENIE-COLU-00206-01	76	LUAD	Primary	COLU-TSACP-V1	Non-Small Cell Lung Cancer	Lung Adenocarcinoma	
836	GENIE-COLU-00207	GENIE-COLU-00207-01	67	LUAD	Primary	COLU-TSACP-V1	Non-Small Cell Lung Cancer	Lung Adenocarcinoma	

ONCOTREE_CODE	SAMPLE_TYPE	count
<chr>	<chr>	<int>
COAD	Metastasis	2179
COAD	Primary	4089
IDC	Metastasis	2357
IDC	Primary	5669
LUAD	Metastasis	4951
LUAD	Primary	7582
PAAD	Metastasis	1468
PAAD	Primary	2105
PRAD	Metastasis	1548
PRAD	Primary	2233



Think: Why the result looks "different" with the same dataset?