Yufei_HW2

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

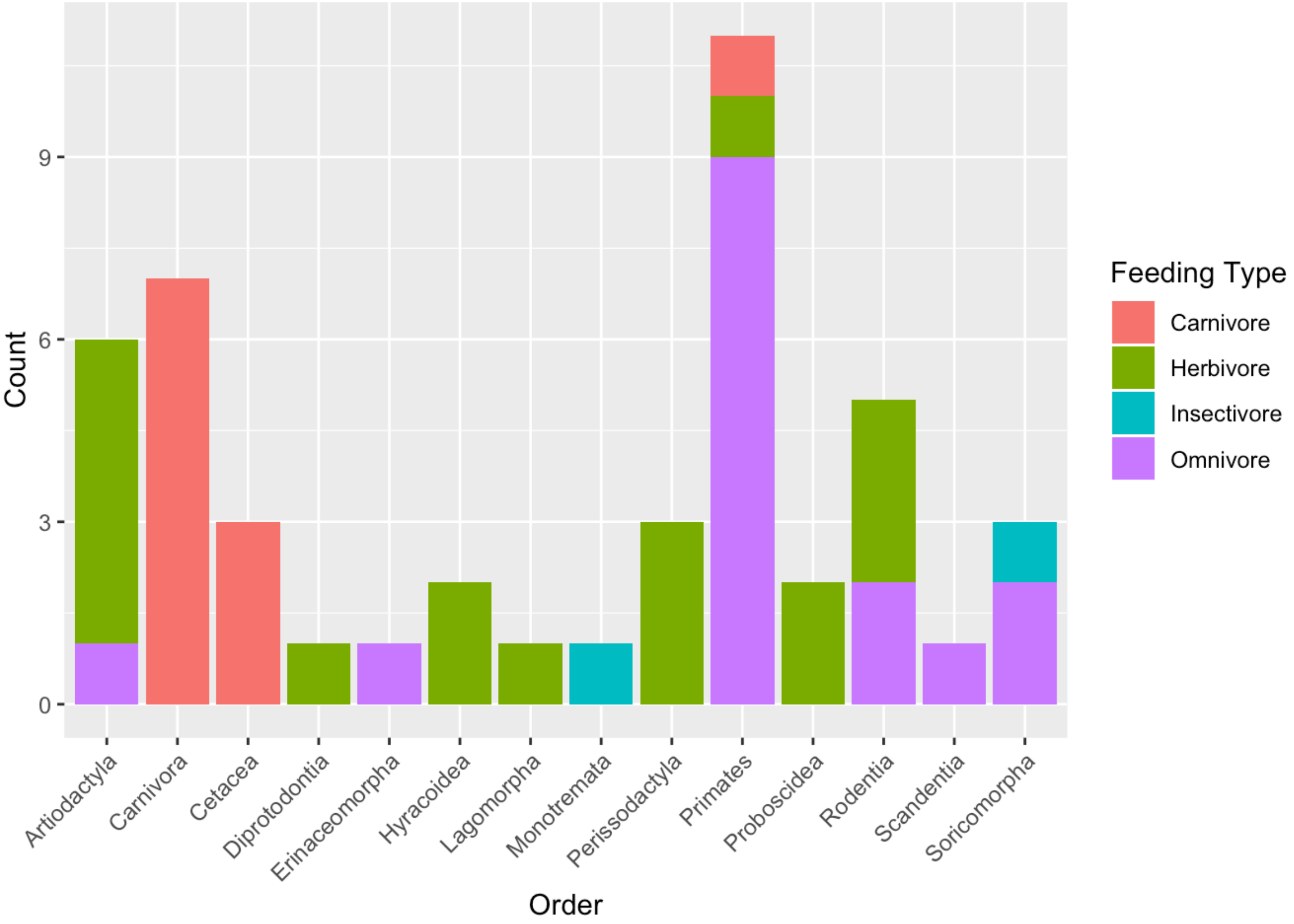
```
#Section A
#Question 1
#loading the libraries
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────── tidyvers
e 1.2.1 ──
```

```
## ✔ ggplot2 3.2.1     ✔ purrr   0.3.3
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   1.0.0     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────── tidyverse_conf
licts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```r
library(dplyr)

#set up the those animals whose awake time over 12 hours
data1<- filter(msleep,msleep$awake>12)

#remove the NA values from feeding types
data2 <- data1[!is.na(data1$vore),]

#rename the feeding tpypes names
data3<- data2 %>% mutate(vore=recode(vore,
                         `carni`="Carnivore",
                         `herbi`="Herbivore",
                        `insecti`="Insectivore",
                        `omni`="Omnivore" ))
#create the targert bar plot
ggplot(data = data3) +
geom_bar(mapping = aes(x = order, fill = vore)) +
scale_x_discrete() +
scale_fill_discrete() +
scale_y_continuous() +
theme(axis.text.x=element_text(angle = 45, hjust = 1))+
labs(fill="Feeding Type")+
xlab("Order")+
ylab("Count")
```

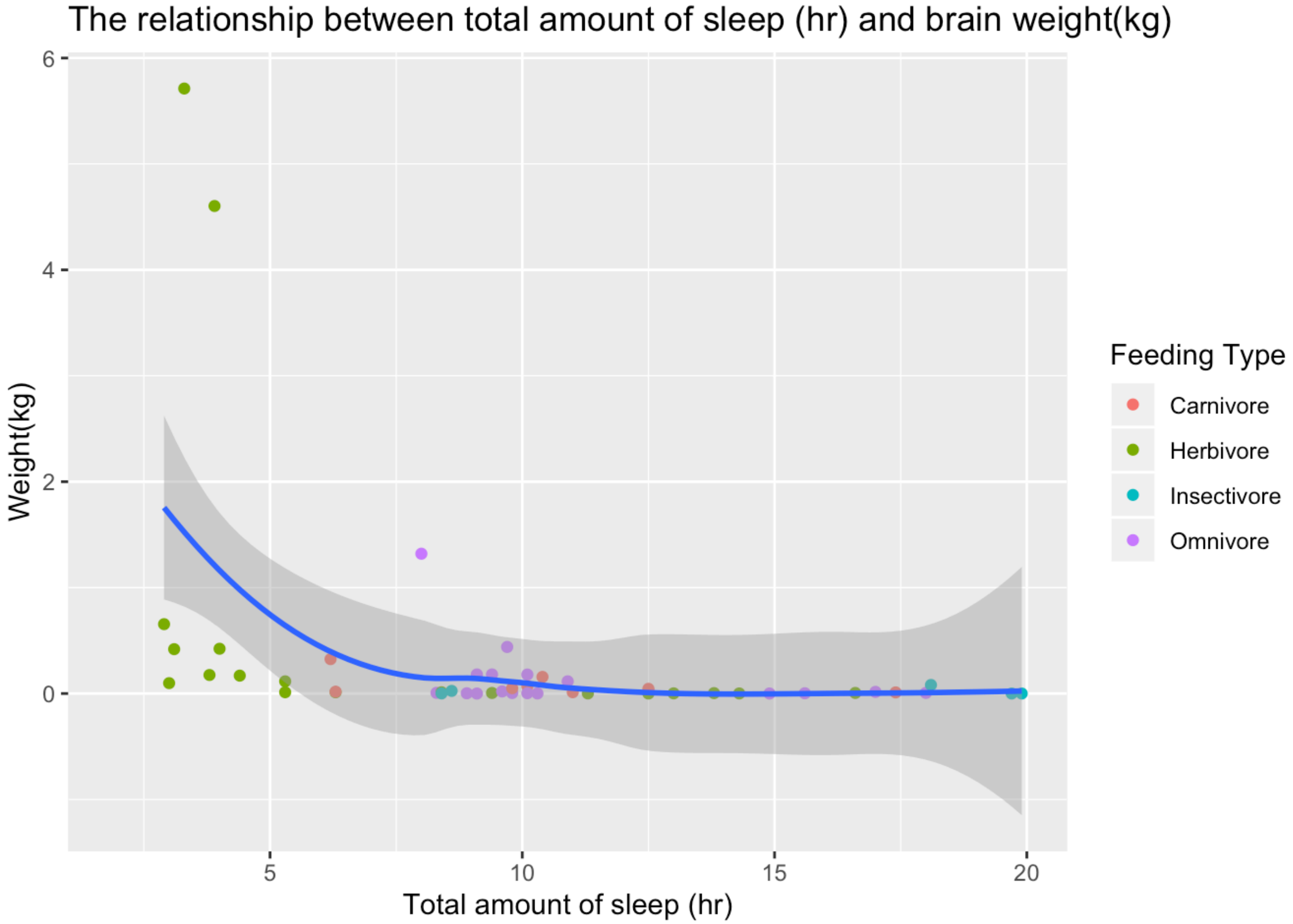

Count
9
6
3
0
Order
Artiodactyla
Carnivora
Cetacea
Diprotodontia
Erinaceomorpha
Hyracoidea
Lagomorpha
Monotremata
Perissodactyla
Primates
Proboscidea
Rodentia
Scandentia
Soricomorpha
Feeding Type
Carnivore
Herbivore
Insectivore
Omnivore

```
#Question 2

#remove the NA values from feeding types
data4 <- msleep[!is.na(msleep$vore),]

#rename the feeding tpypes names
data5<- data4 %>% mutate(vore=recode(vore,
                         `carni`="Carnivore",
                         `herbi`="Herbivore",
                        `insecti`="Insectivore",
                        `omni`="Omnivore" ))

#get the target plot
ggplot(data=data5,mapping=aes(x=sleep_total,y=brainwt))+
  geom_point(mapping=aes(color=vore))+
  geom_smooth()+
  xlab("Total amount of sleep (hr)")+
  ylab("Weight(kg)")+
  ggtitle("The relationship between total amount of sleep (hr) and b
rain weight(kg)")+
  labs(color="Feeding Type")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 25 rows containing non-finite values (stat_smoot
h).
```

```
## Warning: Removed 25 rows containing missing values (geom_point).
```



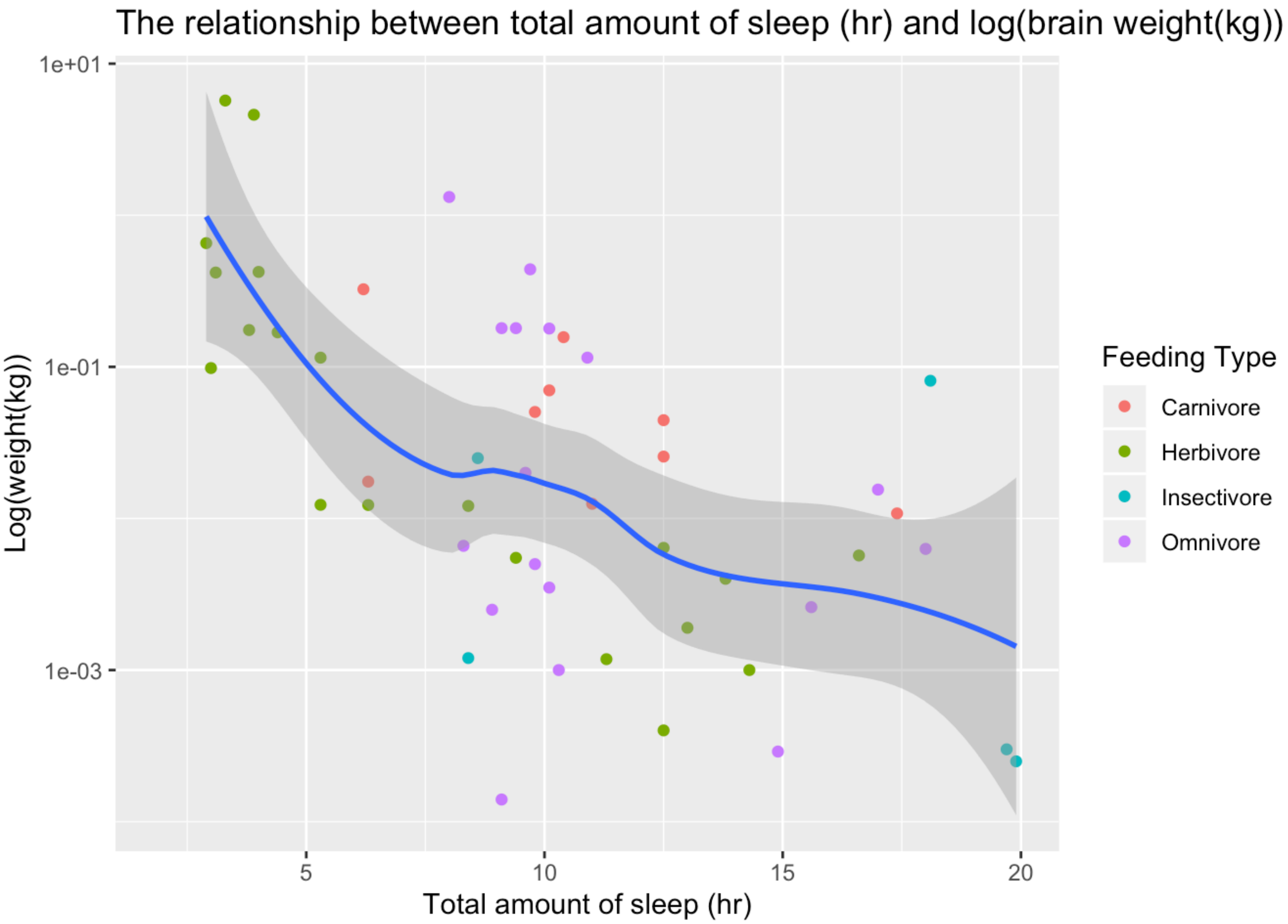
```
#Answer:
#From the plot below, there are a lots of outliers,and most mammals`
weights are relevant small.
#Therefore, it is hard to see the relationship between total slepp h
ours and mammal` body weight
```

```
#Question3

#get the target plot
ggplot(data=data5,mapping=aes(x=sleep_total,y=brainwt))+
  geom_point(mapping=aes(color=vore))+
  geom_smooth()+
  xlab("Total amount of sleep (hr)")+
  ylab("Log(weight(kg))")+
  ggtitle("The relationship between total amount of sleep (hr) and l
og(brain weight(kg))")+
  labs(color="Feeding Type")+
  scale_y_log10()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 25 rows containing non-finite values (stat_smoot
h).
```

```
## Warning: Removed 25 rows containing missing values (geom_point).
```



```
#Answer
#From the plot below,the more the mammal weighs, the less likely the
total hour of sleep.
```

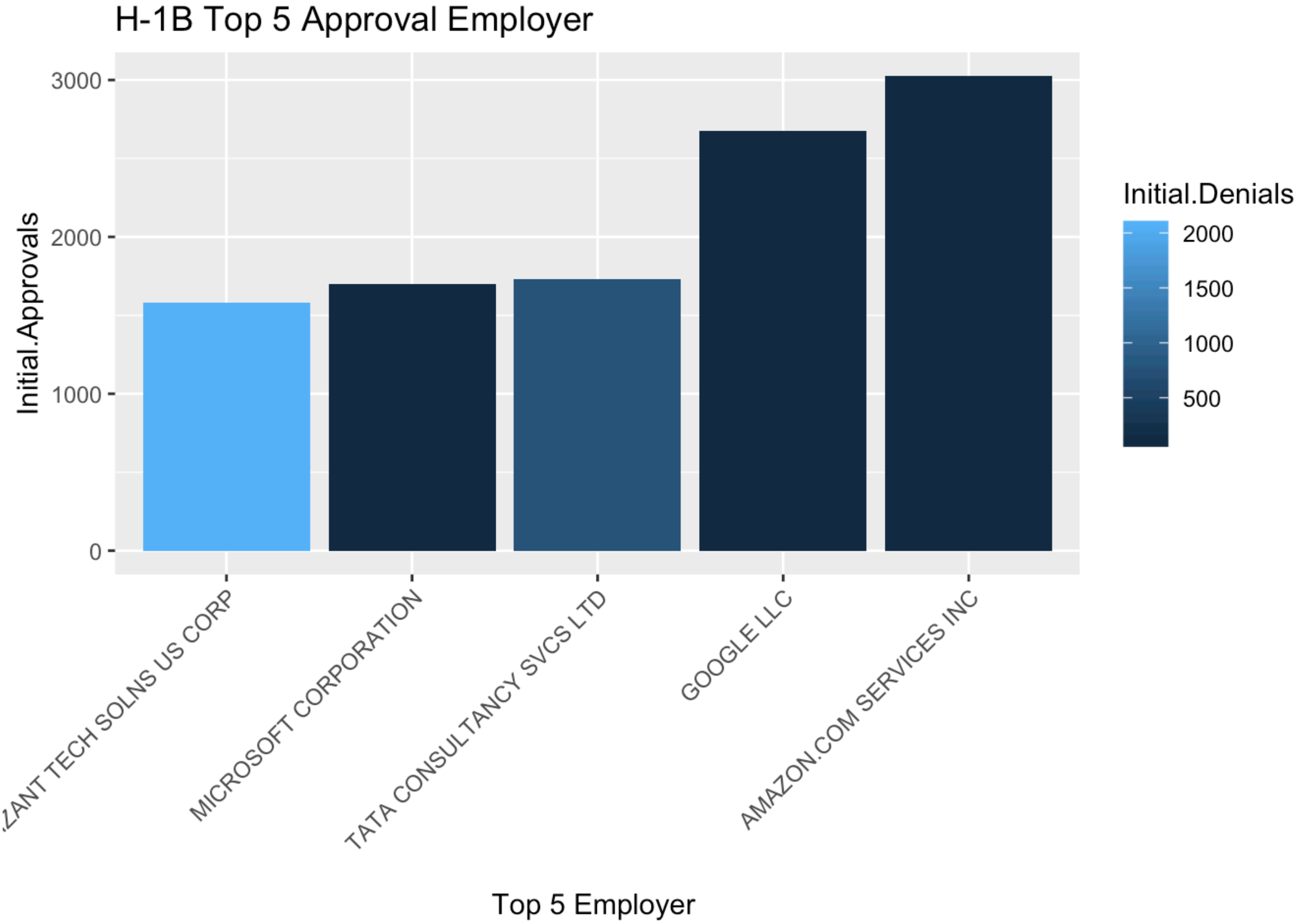
```
#SectionB
#Question1
library(readr)

#loading the dataset
my_data <- read.csv("h1b_datahubexport-2019.csv",stringsAsFactors =
FALSE)

#covert the data type of columns in data frame
#my_data[, c(3:6)] <- sapply(my_data[, c(3:6)], as.numeric)
rang_rows <- 3:6
my_data[,rang_rows] <- lapply(my_data[,rang_rows],function(my_data)
  {as.numeric(gsub(",","",my_data)) })

# check the data type
sapply(my_data,mode)
```

```
##          Fiscal.Year             Employer    Initial.Approvals
##            "numeric"          "character"            "numeric"
##      Initial.Denials Continuing.Approvals   Continuing.Denials
##            "numeric"            "numeric"            "numeric"
##                NAICS               Tax.ID                State
##            "numeric"            "numeric"          "character"
##                 City                  ZIP
##          "character"            "numeric"
```

```
#get the top 5 employers which have the most cases of initial approv
ed H-1B
my_data1 <- head(arrange(my_data,desc(Initial.Approvals)),n=5)
# get the new top 5 dataframe
my_data2 <- my_data1[,c(2,3,4,5,6)]
my_data2
```

```
##                        Employer Initial.Approvals Initial.Denials
## 1      AMAZON.COM SERVICES INC              3026             122
## 2                   GOOGLE LLC              2678             104
## 3    TATA CONSULTANCY SVCS LTD              1733             763
## 4        MICROSOFT CORPORATION              1701             109
## 5 COGNIZANT TECH SOLNS US CORP              1580            2060
##   Continuing.Approvals Continuing.Denials
## 1                 4186                133
## 2                 3333                 53
## 3                 5859               1376
## 4                 3560                 66
## 5                11783               3910
```

```
#plot a bar chart of Employer versus Initial approvals
ggplot(data=my_data2)+
geom_bar(mapping=aes(x=reorder(Employer,Initial.Approvals),y=Initial
.Approvals,fill=Initial.Denials),stat="identity")+
scale_x_discrete() +
scale_y_continuous() +
theme(axis.text.x=element_text(angle = 45, hjust = 1))+
xlab("Top 5 Employer ")+
ggtitle("H-1B Top 5 Approval Employer")
```



```
#what do you notice based on the plot?
#Answer
#The company(Amazon) that gets the highest initial approvals of H1-B sponserships
# also gets the lowest initial denials of the same sponsership.
```

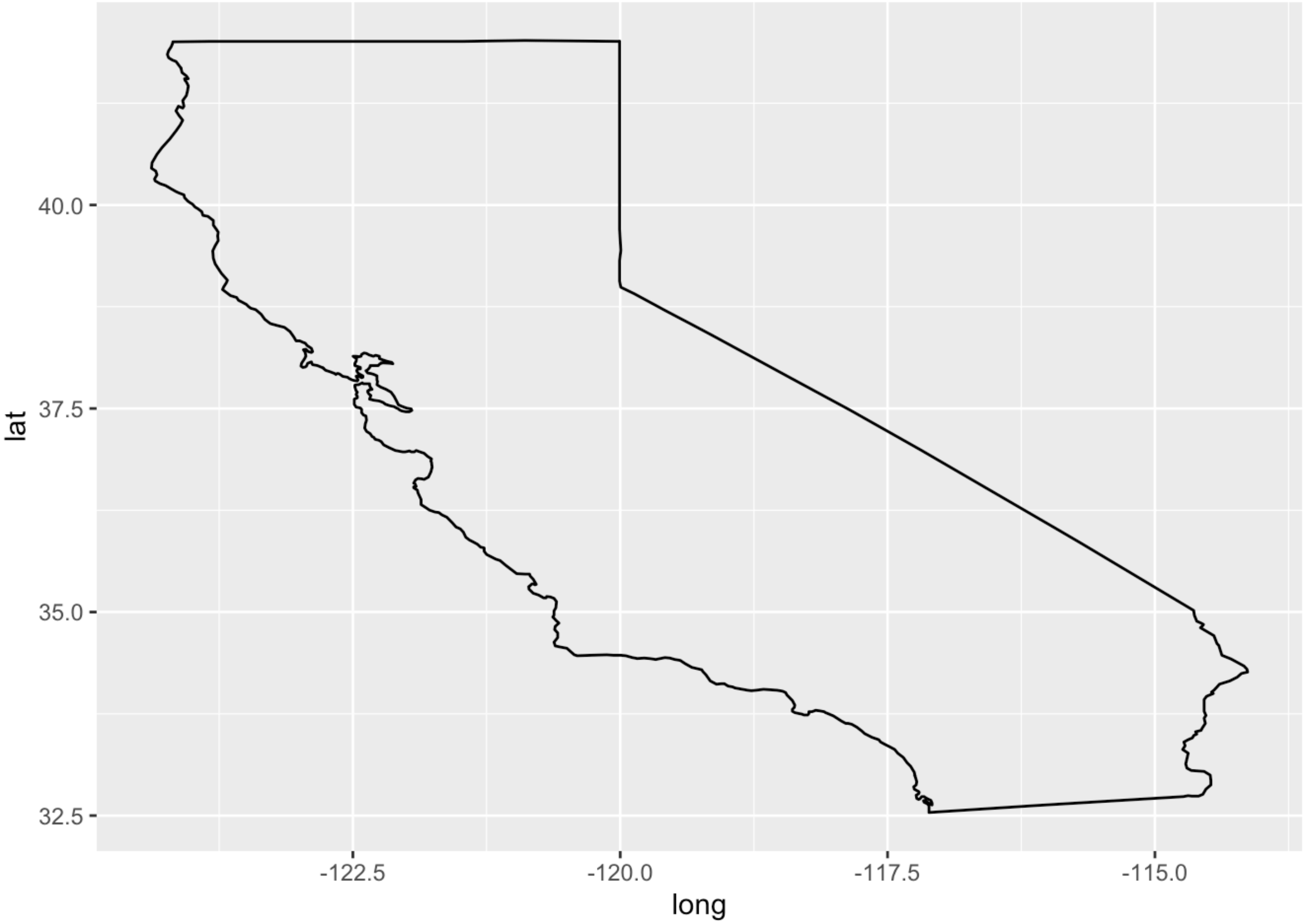
```
#Question 2
library(readr)
#loading the geocode dataset
result<- read.csv("us-zip-code-latitude-and-longitude.csv",sep=";",header=TRUE)

#prepare for the join two tables
names(result)[names(result) == "Zip"] <- "ZIP"
result[,2] = toupper(result[,2])
result1 <- select(result,ZIP,City,State,Latitude,Longitude)
result2 <-my_data[,c(2,3,4,5,6,9,10,11)]

#join two tables by three common columns
new_dataset <-merge(result1,result2,by=c('ZIP','State','City'))

#use the mutate function to add a new column
new_dataset1 <- mutate(new_dataset,Prop=Initial.Denials/Initial.Approvals)
```

```
#Question3
library(ggplot2)
library(maps)
```

```
##
## Attaching package: 'maps'
```

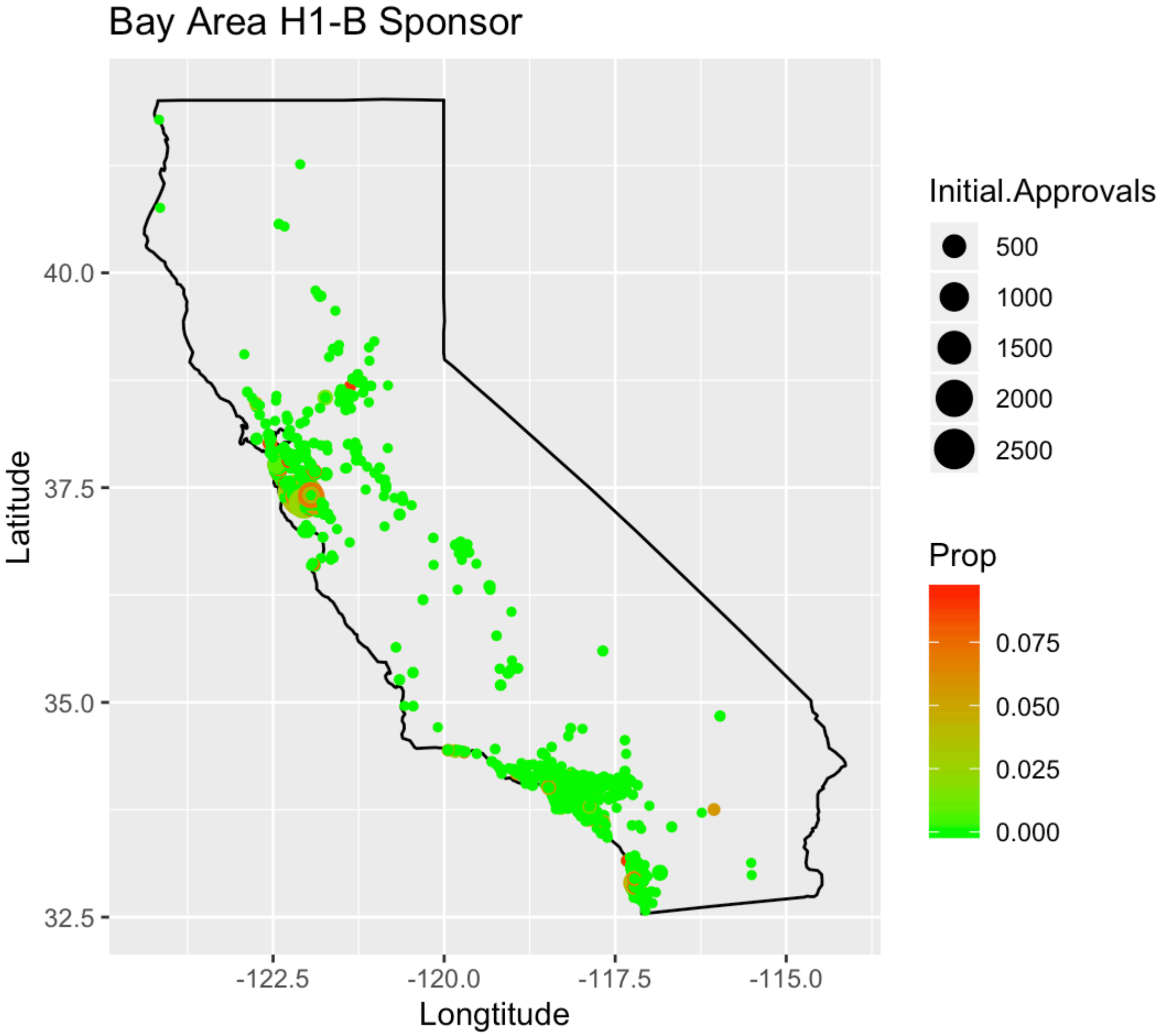
```
## The following object is masked from 'package:purrr':
##
##     map
```

```
#get the dataset of longitudes and latitudes of CA
ca_map <- map_data("state",region="California")
#plot the CA map
a <- ggplot(ca_map) +
geom_polygon(mapping = aes(x = long, y = lat,group=group),color = "black",fill=NA)
a
```



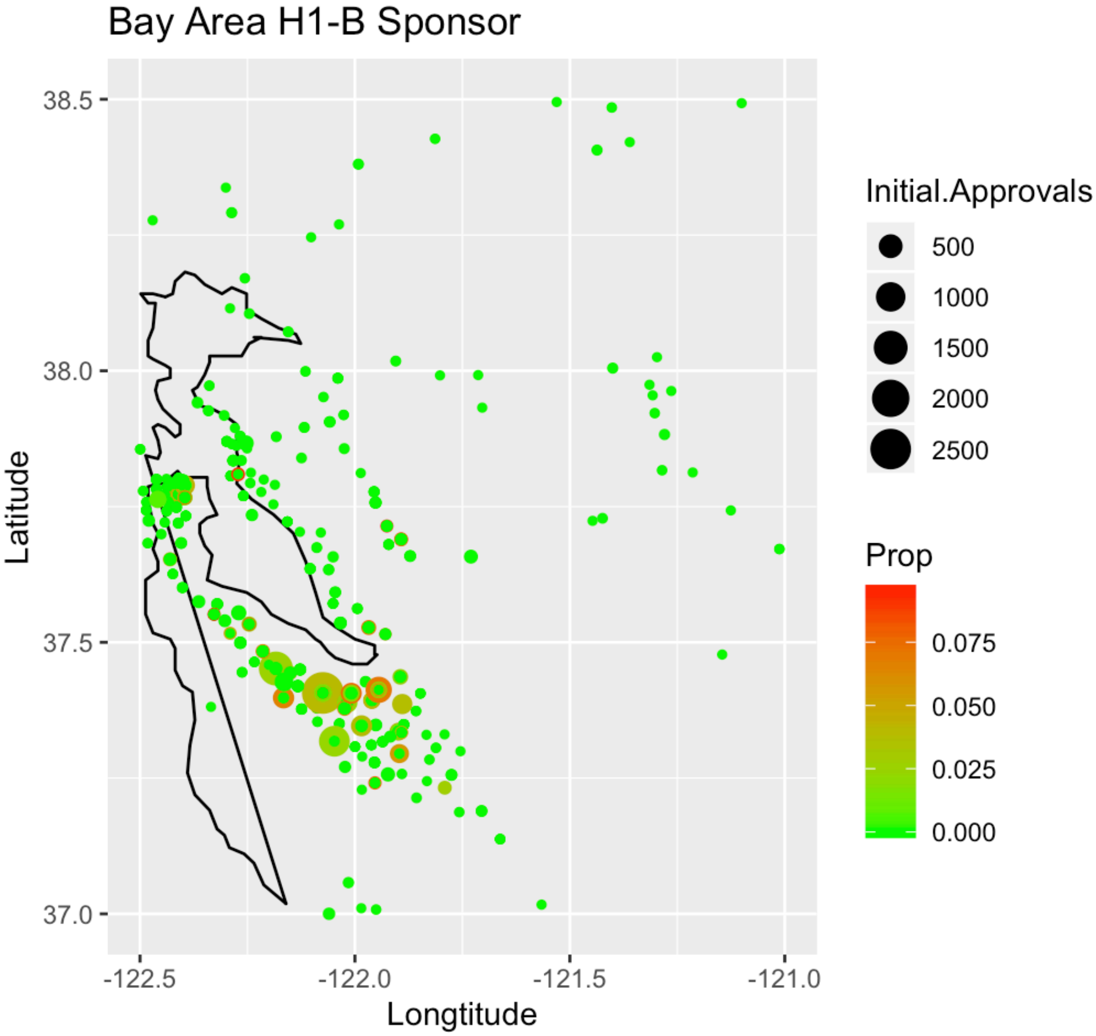
```
#filter the H1-B dataset
names(new_dataset1)[names(new_dataset1) == "Latitude"] <- "lat"
names(new_dataset1)[names(new_dataset1) == "Longitude"] <- "long"
new_dataset2 <- new_dataset1[new_dataset1$Prop < 0.1 & new_dataset1$
State=="CA",]

#get the target plot before putting x and y lim
 b <- a+geom_point(data=new_dataset2,mapping = aes(x = long, y = lat
, color = Prop,size=Initial.Approvals))+
 coord_quickmap()+
  xlab("Longtitude")+
 ylab("Latitude")+
  ggtitle("Bay Area H1-B Sponsor")+
   scale_color_gradient(low="green",high="red")
b
```

```
## Warning: Removed 4119 rows containing missing values (geom_point)
.
```



```
#get the target plot after putting the x and y lim
 c <- a+geom_point(data=new_dataset2,mapping = aes(x = long, y = lat
, color = Prop,size=Initial.Approvals))+
 coord_quickmap()+
  xlab("Longtitude")+
 ylab("Latitude")+
  ggtitle("Bay Area H1-B Sponsor")+
   scale_color_gradient(low="green",high="red")+
 xlim(c(-122.5,-121))+
  ylim(c(37,38.5))
 c
```

```
## Warning: Removed 6563 rows containing missing values (geom_point)
.
```


Bay Area H1-B Sponsor
38.5
38.0
37.5
37.0
Latitude
-122.5
-122.0
-121.5
-121.0
Longtitude
Initial.Approvals
500
1000
1500
2000
2500
Prop
0.075
0.050
0.025
0.000