Home-Work(2)

```
                                   HomeWork-2



                                    Section A
```

Only use ggplot2 for plotting This section is for testing your ggplot2, and data exploration skills. Dataset msleep from ggplot2 package will be using through this section. Use ? to check the documentation of msleep.

Problem 1 We are interested in those animals whose awake time over 12 hours. Create a bar chart as the following figure. Remove the NA values from feeding types: carnivore, omnivore, insectivore and herbivore. hints: You may adjust the angel of x-axis label by using theme(axis.text.x=element_text()), and the legend labels by using scale_fill_discrete().

```
install.packages('readr')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages('tidyverse')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
library('ggplot2')
library('readr')
library('tidyverse')
```

```
## ── Attaching packages ───────────────────────────────────────────────────────
────────── tidyverse 1.3.0 ──
```

```
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.0     ✓ stringr 1.4.0
## ✓ purrr   0.3.3     ✓ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────────
──── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
data("msleep")
head(msleep)
```

```
## # A tibble: 6 x 11
##   name  genus vore  order conservation sleep_total sleep_rem sleep_cycle
##   <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl>
## 1 Chee… Acin… carni Carn… lc                  12.1      NA        NA
## 2 Owl … Aotus omni  Prim… <NA>                17         1.8      NA
## 3 Moun… Aplo… herbi Rode… nt                  14.4       2.4      NA
## 4 Grea… Blar… omni  Sori… lc                  14.9       2.3       0.133
## 5 Cow   Bos   herbi Arti… domesticated         4         0.7       0.667
## 6 Thre… Brad… herbi Pilo… <NA>                14.4       2.2       0.767
## # … with 3 more variables: awake <dbl>, brainwt <dbl>, bodywt <dbl>
```

```
?msleep
```

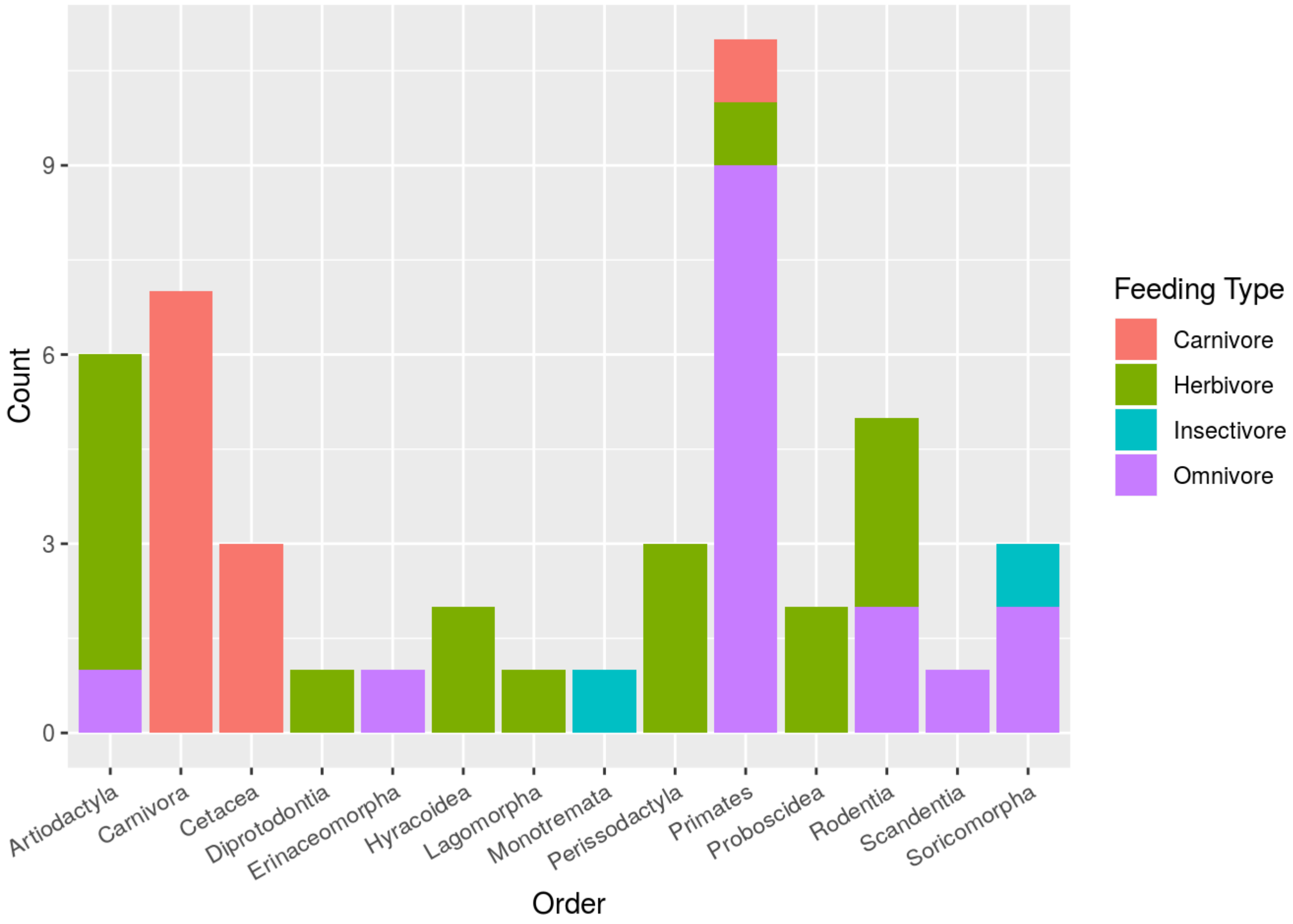
```
##### Dropping NA values
m<-c('vore','order','awake')
new<-msleep[m]
new<-new[new$awake>12,]
new<-new%>%drop_na(vore)
sum(is.na(new$vore))
```

```
## [1] 0
```

```
head(new)
```

```
## # A tibble: 6 x 3
##   vore  order        awake
##   <chr> <chr>        <dbl>
## 1 herbi Artiodactyla  20  
## 2 carni Carnivora     15.3
## 3 carni Carnivora     13.9
## 4 herbi Artiodactyla  21  
## 5 herbi Artiodactyla  18.7
## 6 herbi Rodentia      14.6
```

```
##### Plotting the bar graph

ggplot(data=new)+geom_bar(mapping=aes(order,fill=vore))+scale_fill_discrete(name='Feeding Type',
labels=c('Carnivore','Herbivore','Insectivore','Omnivore'))+theme(axis.text.x = element_text(ang
le=30,hjust=1))+xlab('Order')+ylab('Count')
```



Problem 2 We would like to investigate how’s the relationship between total amount of sleep (hr) and brain weight(kg) among feeding types: carnivore, omnivore, insectivore and herbivore. Plot total amount of sleep (hr) versus brain weight (kg), applying color mapping on the feeding types(vore). Remove the NA group from feeding types. Include a smoothing line on the plot. What do you notice in the plot?

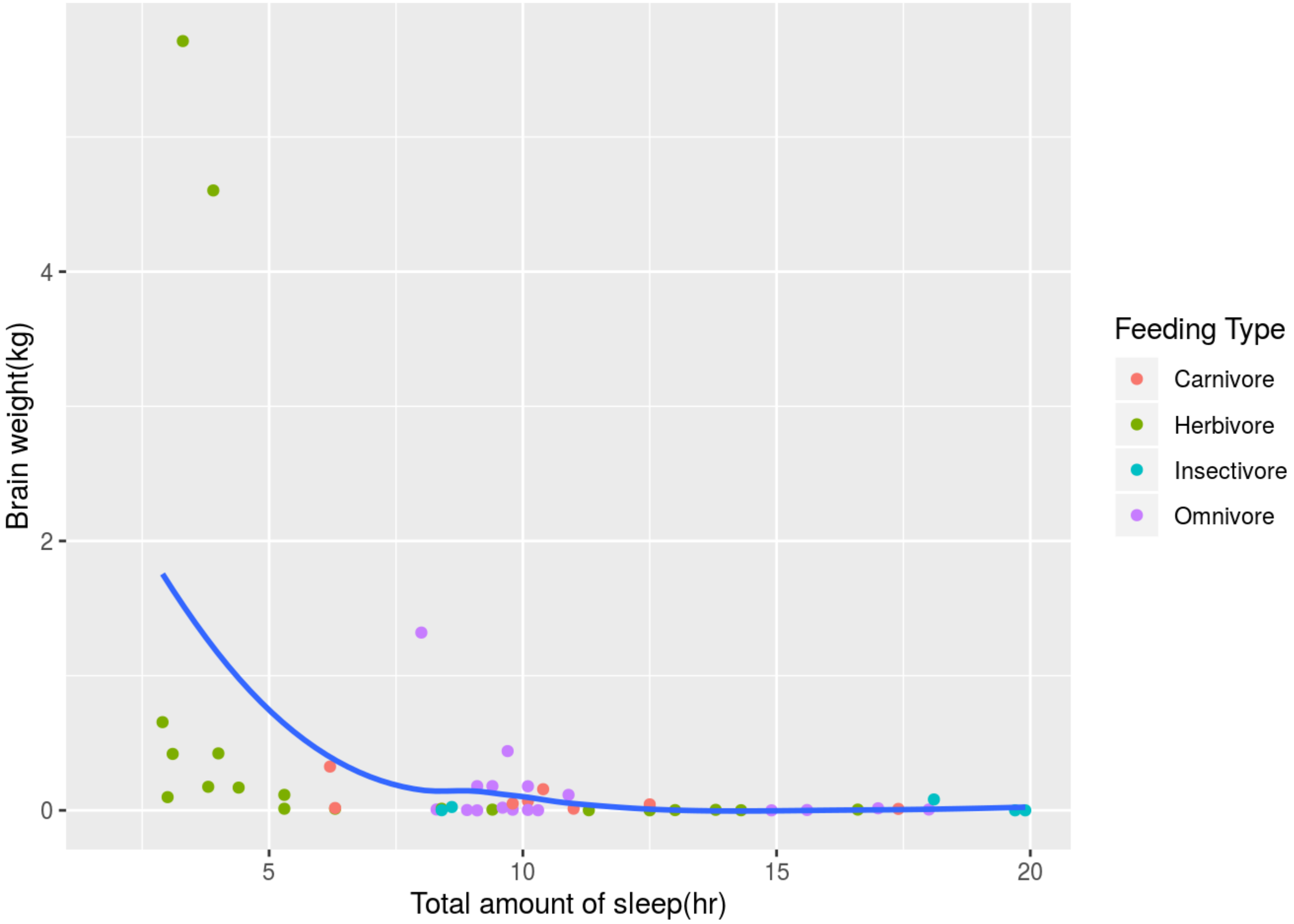
```
m<-c('vore','brainwt','sleep_total')
new<-msleep[m]
new<-new%>%drop_na(vore)
sum(is.na(new$vore))
```

```
## [1] 0
```

```
head(new)
```

```
## # A tibble: 6 x 3
##   vore   brainwt sleep_total
##   <chr>    <dbl>       <dbl>
## 1 carni NA              12.1
## 2 omni   0.0155         17
## 3 herbi NA              14.4
## 4 omni   0.00029        14.9
## 5 herbi  0.423           4
## 6 herbi NA              14.4
```

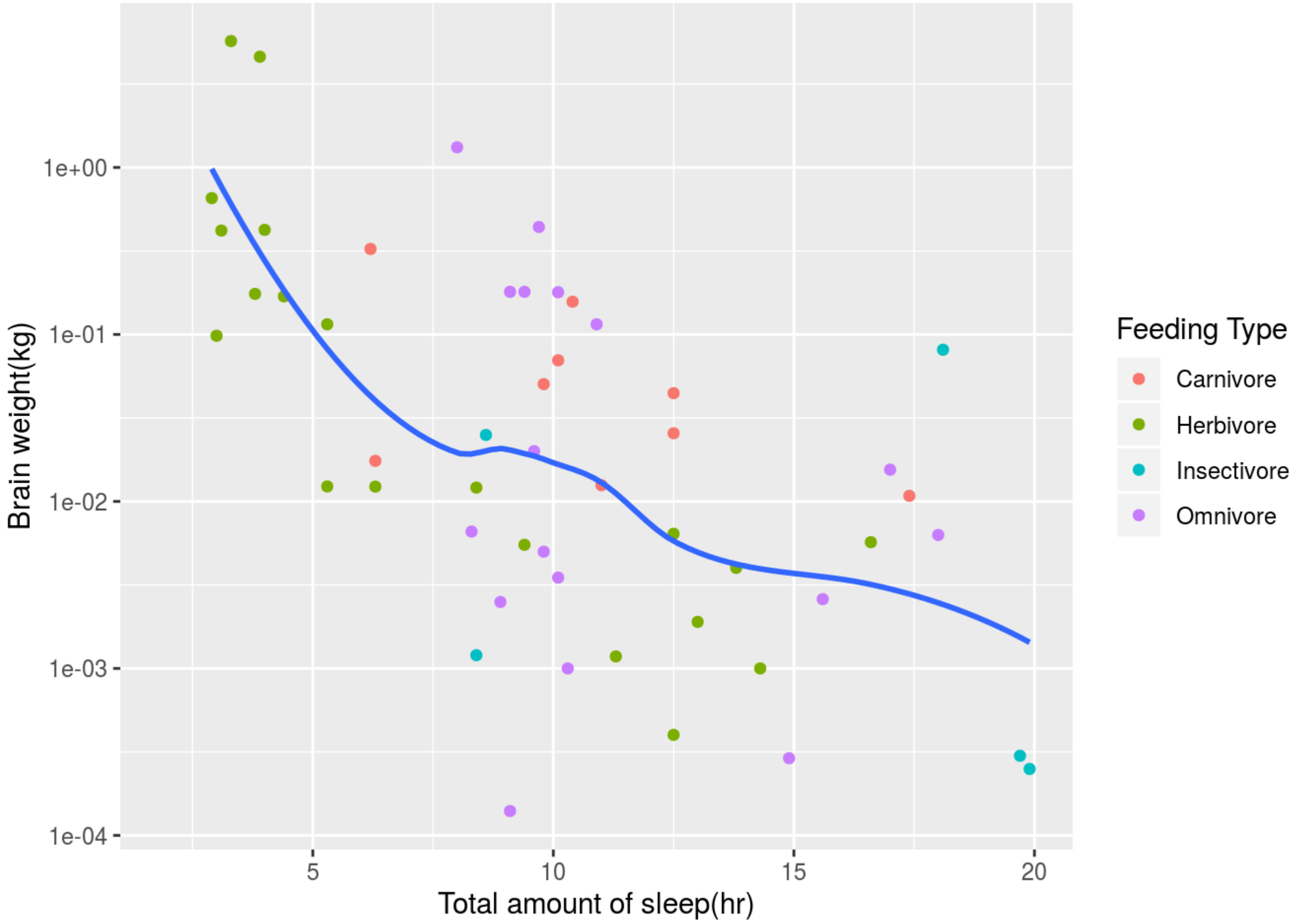
```
##### Plotting the graph
ggplot(data=new,aes(x=sleep_total,y=brainwt))+geom_point(aes(color=vore))+geom_smooth(se=FALSE)+
xlab('Total amount of sleep(hr)')+ylab('Brain weight(kg)')+scale_color_discrete(name='Feeding Ty
pe',labels=c('Carnivore','Herbivore','Insectivore','Omnivore'))
```



Interpretation: The brain weight of the animals are distributed in the range of 0.001 to 0.4, the smoothing lines helps us to understand the relationship between brain weight and total amount of sleep, the smoothing line linearly decreases and then it reaches to a constant. Most of the animals have similar brain weight with sleeping hours around 10 hours. It is difficult to interpret what is the relationship, because most of the data are skewed aroung 10 hours. There is also two outliers in the plot.

Problem 3 Still working on the above plot. Apply log transformation on the brain weight Brain Weight (Kg), Log, what do you observe in the plot?

```
##### Plotting using log transformation

ggplot(data=new,aes(x=sleep_total,y=brainwt))+geom_point(aes(color=vore))+geom_smooth(se=FALSE)+
scale_y_log10()+xlab('Total amount of sleep(hr)')+ylab('Brain weight(kg)')+scale_color_discrete
(name='Feeding Type',labels=c('Carnivore','Herbivore','Insectivore','Omnivore'))
```



Interpretation: When log transformation is applied to Brain weight the data's are normalised and the smoothing line indicates that the relationship between two variables decreases linearly. After this transformation it is easy to interpret the relationship between total amount of sleep vs brain weight, with Herbivore having maximum brain weight with lesser sleeping hours, whereas the insectivore have higher sleeping hours with lesser brain weight, which can be interpreted easily in comparision with older plot.

Section B

Only use ggplot2 for plotting Section B uses FY 2019 H-1B Employer Data from U.S. Citizenship and Immigration Services. Download FY2019 H-1B data from: https://www.uscis.gov/tools/reports-studies/h-1b-employer-data-hub-files (https://www.uscis.gov/tools/reports-studies/h-1b-employer-data-hub-files) To read the data manual: https://www.uscis.gov/tools/reports-studies/understanding-our-h-1b-employerdata-hub (https://www.uscis.gov/tools/reports-studies/understanding-our-h-1b-employerdata-hub) The H-1B is a visa in the United States under the Immigration and Nationality Act, section 101(a)(15)(H) that allows U.S. employers to temporarily employ foreign workers in specialty occupations. A specialty occupation requires the application of specialized knowledge and a bachelor's degree or the equivalent of work experience. Use read.csv() to import the dataset to R.

Problem 1 Import the H-1B data. • You may notice the data types of "Initial.Approvals", "Initial.Denials", "Continuing.Approvals", and "Continuing.Denials" are wrong. We need to convert them into numerical columns. • Return a data frame containing the top 5 employers which have the most cases of initial approved H-1B. This data frame should have the columns: employer, initial approvals, initial denials, continuing approvals, and continuing denials. Show the top 5 data frame. • Plot a bar chart of Employer versus Initial approvals, maping Initial Denials as fill, what do you notice based on the plot?

```
##### Importing the data
h1b<-read.csv('h1b_datahubexport-2019.csv',sep=",",stringsAsFactors = FALSE)
head(h1b)
```

```
##   Fiscal.Year                           Employer Initial.Approvals
## 1        2019 SOUTHERN CARPET HARDWOOD & TILE IN                 1
## 2        2019                  UAB HEALTH SYSTEM                 0
## 3        2019       BIRMINGHAM VA MEDICAL CENTER                 0
## 4        2019                GESTAMP ALABAMA LLC                 1
## 5        2019              ARKANSAS HEALTH GROUP                 0
## 6        2019     UNIV OF ARKANSAS AT MONTICELLO                 1
##   Initial.Denials Continuing.Approvals Continuing.Denials NAICS Tax.ID
## 1               0                    0                  0    23     NA
## 2               0                    0                  1    56     NA
## 3               0                    1                  0    62     NA
## 4               0                    0                  0    33     NA
## 5               0                    1                  0    62     NA
## 6               0                    0                  0    61     NA
##   State        City   ZIP
## 1    AL  BIRMINGHAM 35209
## 2    AL  BIRMINGHAM 35233
## 3    AL  BIRMINGHAM 35233
## 4    AL    MC CALLA 35111
## 5    AR LITTLE ROCK 72211
## 6    AR  MONTICELLO 71656
```

```
##### Transforming the data to numeric

h1b<-transform(h1b,Initial.Approvals=as.numeric(gsub(",","",Initial.Approvals)),Initial.Denials=
as.numeric(gsub(",","",Initial.Denials)),Continuing.Approvals=as.numeric(gsub(",","",Continuing.
Approvals)),
               Continuing.Denials=as.numeric(gsub(",","",Continuing.Denials)))

head(h1b)
```

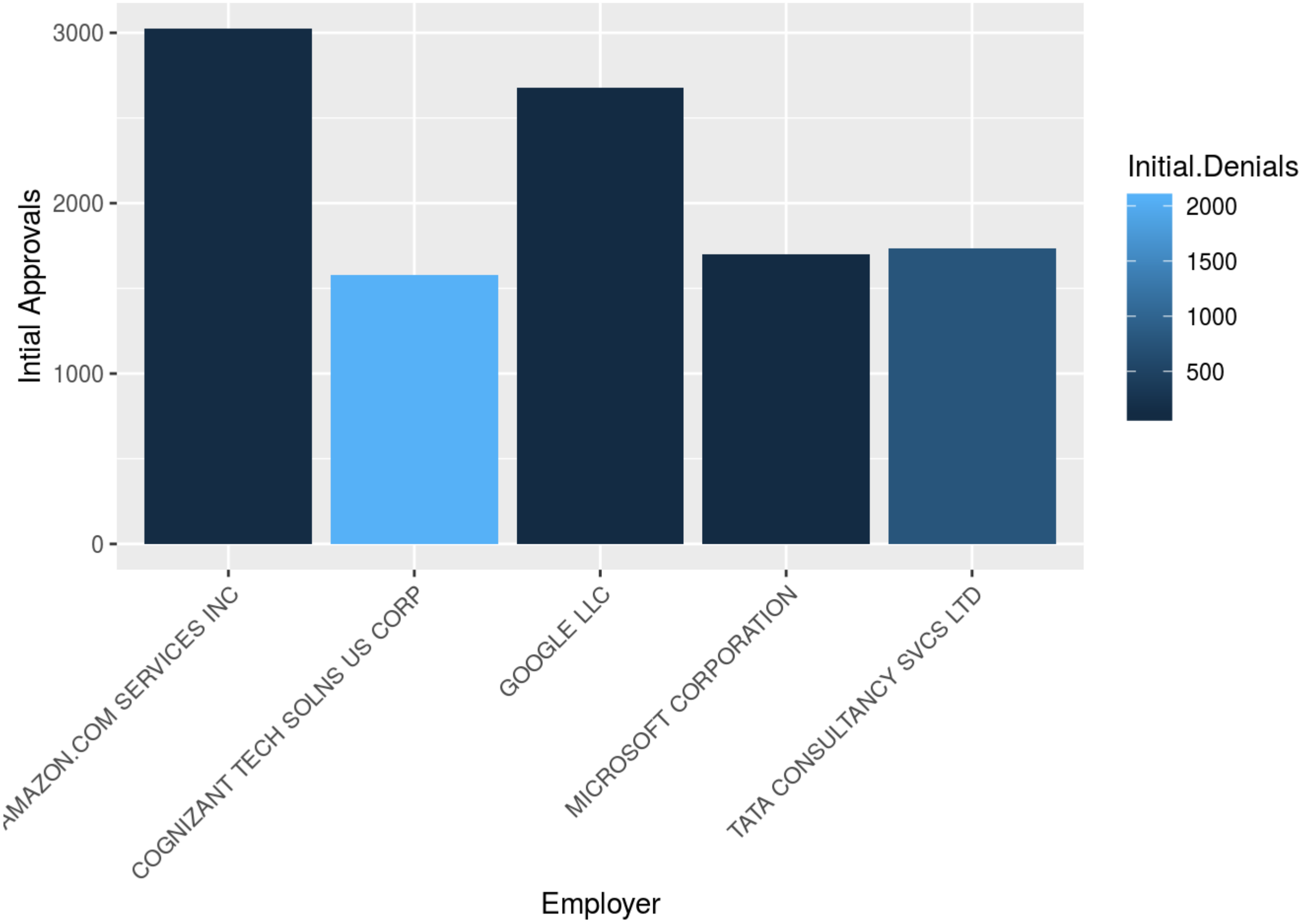
```
##   Fiscal.Year                           Employer Initial.Approvals
## 1        2019 SOUTHERN CARPET HARDWOOD & TILE IN                 1
## 2        2019                  UAB HEALTH SYSTEM                 0
## 3        2019       BIRMINGHAM VA MEDICAL CENTER                 0
## 4        2019                GESTAMP ALABAMA LLC                 1
## 5        2019              ARKANSAS HEALTH GROUP                 0
## 6        2019     UNIV OF ARKANSAS AT MONTICELLO                 1
##   Initial.Denials Continuing.Approvals Continuing.Denials NAICS Tax.ID
## 1               0                    0                  0    23     NA
## 2               0                    0                  1    56     NA
## 3               0                    1                  0    62     NA
## 4               0                    0                  0    33     NA
## 5               0                    1                  0    62     NA
## 6               0                    0                  0    61     NA
##   State        City   ZIP
## 1    AL  BIRMINGHAM 35209
## 2    AL  BIRMINGHAM 35233
## 3    AL  BIRMINGHAM 35233
## 4    AL    MC CALLA 35111
## 5    AR LITTLE ROCK 72211
## 6    AR  MONTICELLO 71656
```

Return a data frame containing the top 5 employers which have the most cases of initial approved H-1B. This data frame should have the columns: employer, initial approvals, initial denials, continuing approvals, and continuing denials. Show the top 5 data frame.

```
##### Subsetting the data with top approval of H1B based on Initial Approvals

m<-(c('Employer','Initial.Approvals','Initial.Denials','Continuing.Approvals','Continuing.Denial
s'))

h1b_new<-h1b[m]

h1b1<-h1b_new[order(h1b_new$Initial.Approvals,decreasing=T)[1:5],]
h1b1
```

```
##                           Employer Initial.Approvals Initial.Denials
## 51300      AMAZON.COM SERVICES INC              3026             122
## 47554                   GOOGLE LLC              2678             104
## 59158    TATA CONSULTANCY SVCS LTD              1733             763
## 55790        MICROSOFT CORPORATION              1701             109
## 5060  COGNIZANT TECH SOLNS US CORP              1580            2060
##       Continuing.Approvals Continuing.Denials
## 51300                 4186                133
## 47554                 3333                 53
## 59158                 5859               1376
## 55790                 3560                 66
## 5060                 11783               3910
```

```
##### Plotting a bar plot according to approvals:
ggplot(data=h1b1,aes(x=Employer,y=Initial.Approvals))+geom_bar(stat='identity',mapping=aes(fill=
Initial.Denials))+theme(axis.text.x=element_text(angle=45,hjust=1))+xlab('Employer')+ylab('Intia
l Approvals')
```



Problem 2 Download geocode data https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/ (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/) export/?location=3,43.25174,-106.27166&basemap=jawg.streets. • Join H-1B data table with geocode data table by State and Zip columns. • This new data frame should include columns: zip, employer, initial approvals, initial denials, continuing approvals, continuing denials, state, city, longitude, and latitude. • Insert a new column prop into this new data frame by the formula: inital denial/initial approval

```
##### Importing the data

geocode<-read.csv('us-zip-code-latitude-and-longitude.csv',sep=';')
head(geocode)
```

```
##     Zip      City State Latitude Longitude Timezone
## 1 71937      Cove    AR 34.39848 -94.39398       -6
## 2 72044  Edgemont    AR 35.62435 -92.16056       -6
## 3 56171  Sherburn    MN 43.66085 -94.74357       -6
## 4 49430    Lamont    MI 43.01034 -85.89754       -5
## 5 52585  Richland    IA 41.19413 -91.98027       -6
## 6 47520 Cannelton    IN 37.93431 -86.67821       -5
##   Daylight.savings.time.flag            geopoint
## 1                          1 34.398483,-94.39398
## 2                          1 35.624351,-92.16056
## 3                          1 43.660847,-94.74357
## 4                          1 43.010337,-85.89754
## 5                          1 41.194129,-91.98027
## 6                          0 37.934311,-86.67821
```

```
##### Joining the table

h1b_new<-merge(h1b,geocode,by.x=c('ZIP','State'),by.y=c('Zip','State'))

head(h1b_new)
```

```
##     ZIP State Fiscal.Year                           Employer
## 1 10001    NY        2019                HAYMARKET MEDIA INC
## 2 10001    NY        2019                SHINAN BANK AMERICA
## 3 10001    NY        2019    BISLEY INC DBA BISLEY N AMERICA
## 4 10001    NY        2019                     TRIALSPARK INC
## 5 10001    NY        2019                       33ACROSS INC
## 6 10001    NY        2019 ANIKA PHARMACY CORP DBA LORVEN PHA
##   Initial.Approvals Initial.Denials Continuing.Approvals
## 1                 0               1                    0
## 2                 0               0                    1
## 3                 0               0                    1
## 4                 1               0                    3
## 5                 2               0                    0
## 6                 1               0                    0
##   Continuing.Denials NAICS Tax.ID   City.x   City.y Latitude Longitude
## 1                  0    54   1585 NEW YORK New York 40.75074 -73.99653
## 2                  1    52   1762 NEW YORK New York 40.75074 -73.99653
## 3                  0    23   8497 NEW YORK New York 40.75074 -73.99653
## 4                  1    54   4239 NEW YORK New York 40.75074 -73.99653
## 5                  0    54   3623 NEW YORK New York 40.75074 -73.99653
## 6                  0    44   2948 NEW YORK New York 40.75074 -73.99653
##   Timezone Daylight.savings.time.flag            geopoint
## 1       -5                          1 40.750742,-73.99653
## 2       -5                          1 40.750742,-73.99653
## 3       -5                          1 40.750742,-73.99653
## 4       -5                          1 40.750742,-73.99653
## 5       -5                          1 40.750742,-73.99653
## 6       -5                          1 40.750742,-73.99653
```

```
##### Subsetting the dataframe
m<-c('ZIP','Employer','Initial.Approvals','Initial.Denials','Continuing.Approvals','Continuing.D
enials','State','City.x','City.y','Longitude','Latitude')

h1b_new<-h1b_new[m]
head(h1b_new)
```

```
##     ZIP                           Employer Initial.Approvals
## 1 10001                HAYMARKET MEDIA INC                 0
## 2 10001                SHINAN BANK AMERICA                 0
## 3 10001    BISLEY INC DBA BISLEY N AMERICA                 0
## 4 10001                     TRIALSPARK INC                 1
## 5 10001                        33ACROSS INC                 2
## 6 10001 ANIKA PHARMACY CORP DBA LORVEN PHA                 1
##   Initial.Denials Continuing.Approvals Continuing.Denials State   City.x
## 1               1                    0                  0    NY NEW YORK
## 2               0                    1                  1    NY NEW YORK
## 3               0                    1                  0    NY NEW YORK
## 4               0                    3                  1    NY NEW YORK
## 5               0                    0                  0    NY NEW YORK
## 6               0                    0                  0    NY NEW YORK
##     City.y Longitude Latitude
## 1 New York -73.99653 40.75074
## 2 New York -73.99653 40.75074
## 3 New York -73.99653 40.75074
## 4 New York -73.99653 40.75074
## 5 New York -73.99653 40.75074
## 6 New York -73.99653 40.75074
```

```
##### Adding proportion as a new column to the dataframe

h1b_new$prop<-h1b_new$Initial.Denials/h1b_new$Initial.Approvals
head(h1b_new)
```

```
##     ZIP                           Employer Initial.Approvals
## 1 10001                HAYMARKET MEDIA INC                 0
## 2 10001                SHINAN BANK AMERICA                 0
## 3 10001    BISLEY INC DBA BISLEY N AMERICA                 0
## 4 10001                     TRIALSPARK INC                 1
## 5 10001                       33ACROSS INC                 2
## 6 10001 ANIKA PHARMACY CORP DBA LORVEN PHA                 1
##   Initial.Denials Continuing.Approvals Continuing.Denials State   City.x
## 1               1                    0                  0    NY NEW YORK
## 2               0                    1                  1    NY NEW YORK
## 3               0                    1                  0    NY NEW YORK
## 4               0                    3                  1    NY NEW YORK
## 5               0                    0                  0    NY NEW YORK
## 6               0                    0                  0    NY NEW YORK
##     City.y Longitude Latitude prop
## 1 New York -73.99653 40.75074  Inf
## 2 New York -73.99653 40.75074  NaN
## 3 New York -73.99653 40.75074  NaN
## 4 New York -73.99653 40.75074    0
## 5 New York -73.99653 40.75074    0
## 6 New York -73.99653 40.75074    0
```

Problem 3 We are interested in the H-1B cases around Bay Area, California. Create a map of the California, and then adjust the plotting x/y limits to a proper zoom level of Bay Area. Then showing the locations of each employer along with, the prop less than 0.1 (mapped as the color/fill), and the initial approvals (mapped as the size). hints: Install map and mapproj packages, and use the ggplot2::map_data() to draw “California” region of the US.

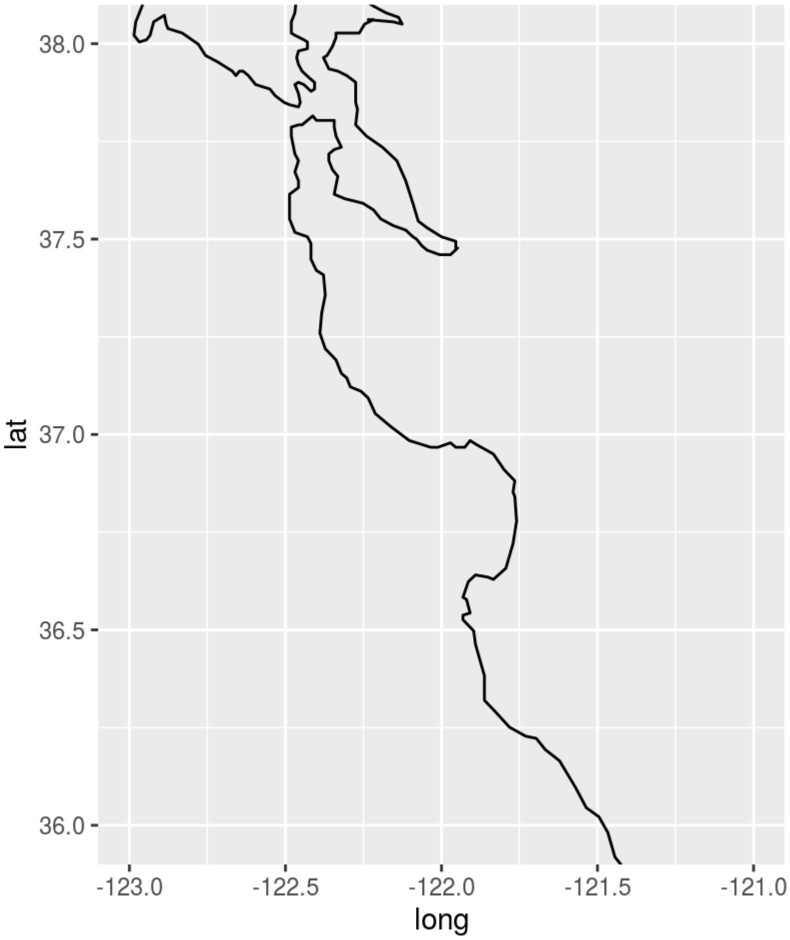
```
install.packages('maps')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages('mapproj')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

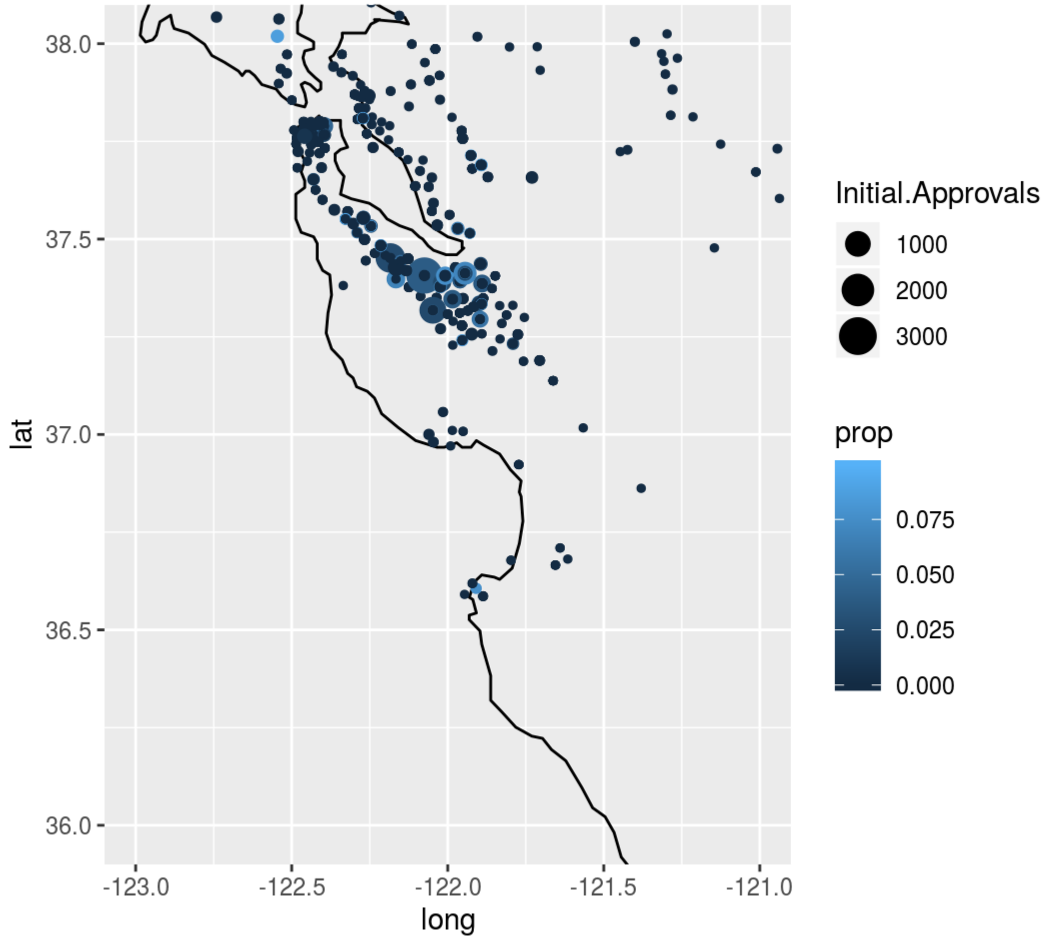
```
##### Reading the california map

cal<- map_data('state',region=c('California'))

ca_base<-ggplot(cal)+geom_polygon(mapping=aes(x=long,y=lat,group=group),fill='NA',color='Black')
+coord_quickmap(xlim = c(-123, -121.0),  ylim = c(36, 38))

ca_base
```



```
##### Finding the proportions less than 0.1

pro<-h1b_new[h1b_new$prop<0.1,]
head(pro)
```

```
##        ZIP                          Employer Initial.Approvals
## NA      NA                              <NA>                NA
## NA.1    NA                              <NA>                NA
## 4    10001                     TRIALSPARK INC                 1
## 5    10001                       33ACROSS INC                 2
## 6    10001 ANIKA PHARMACY CORP DBA LORVEN PHA                 1
## 7    10001                      SHAREBITE INC                 1
##      Initial.Denials Continuing.Approvals Continuing.Denials State
## NA                NA                   NA                 NA  <NA>
## NA.1              NA                   NA                 NA  <NA>
## 4                  0                    3                  1    NY
## 5                  0                    0                  0    NY
## 6                  0                    0                  0    NY
## 7                  0                    0                  0    NY
##        City.x   City.y Longitude Latitude prop
## NA       <NA>     <NA>        NA       NA   NA
## NA.1     <NA>     <NA>        NA       NA   NA
## 4    NEW YORK New York -73.99653 40.75074    0
## 5    NEW YORK New York -73.99653 40.75074    0
## 6    NEW YORK New York -73.99653 40.75074    0
## 7    NEW YORK New York -73.99653 40.75074    0
```

```
##### Plotting the employers and proportions less than 0.1
ca_base+geom_point(data=pro,aes(x=Longitude,y=Latitude,color=prop,size=`Initial.Approvals`))
```


38.0
37.5
37.0
36.5
36.0
lat
-123.0
-122.5
-122.0
-121.5
-121.0
long
Initial.Approvals
1000
2000
3000
prop
0.075
0.050
0.025
0.000