Homework_03_Nikhil

Nikhil Gupta

October 16, 2017

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.

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:

Ques 1

```
##Load the libraries
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
##Read the NBA dataset
nba=read.csv("nba_draft.csv")
##Dataset is not Tidy as the variable "probabilities"" is a combination of
###4 different variables. This implies that each variable does not have its own column
##and consequently each observation does not have its own row
##Use separate to tidy the data, separate probabilites into
##prob_superstar, prob_starter, prob_role_player, prob_bust

nba %>%
  separate(probabilities,into=c("prob_superstar","prob_starter","prob_role_player","prob_bust")
           ,sep=",") ->nba_tidy
##Convert the separated character variable class to numeric variable class
nba_tidy$prob_superstar=as.numeric(nba_tidy$prob_superstar)
nba_tidy$prob_starter=as.numeric(nba_tidy$prob_starter)
nba_tidy$prob_role_player=as.numeric(nba_tidy$prob_role_player)
nba_tidy$prob_bust=as.numeric(nba_tidy$prob_bust)
```

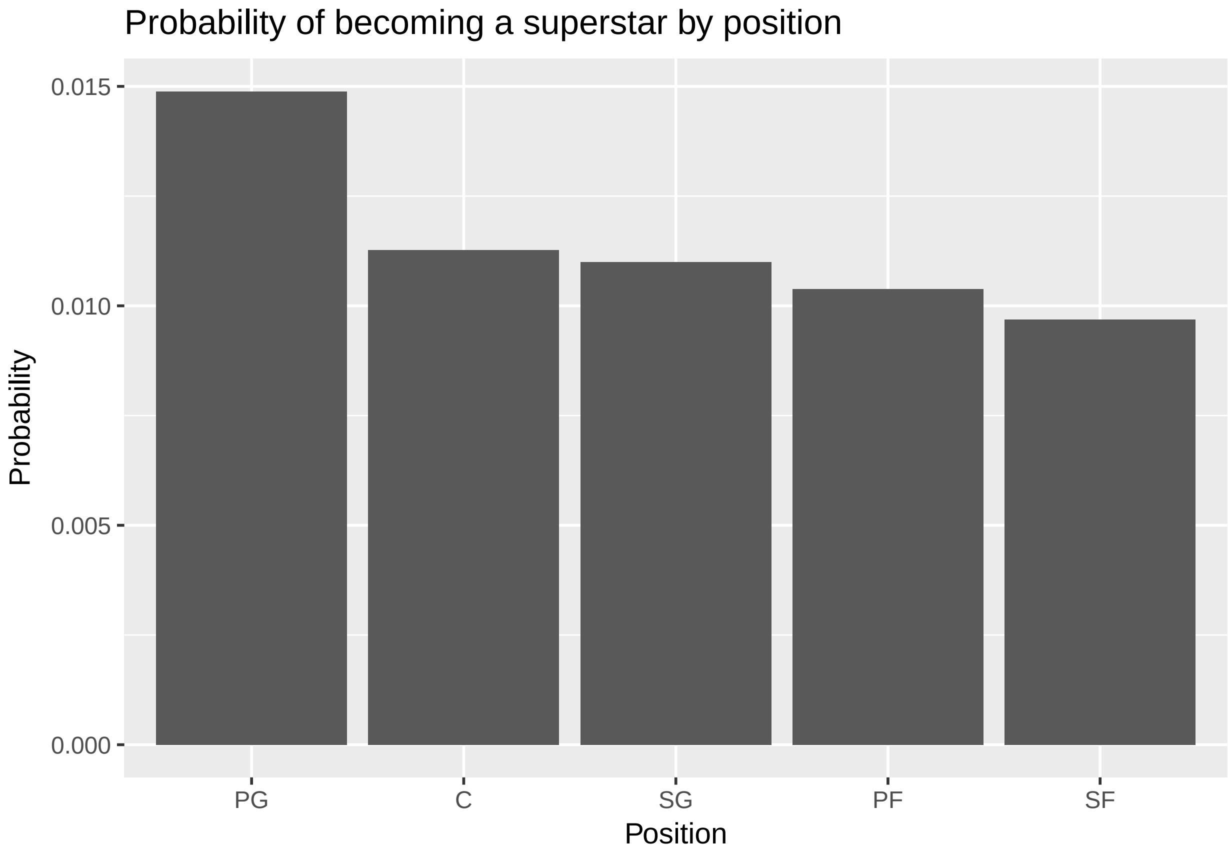
Ques 2

```
paste("The name of the Center with the highest probability of becoming a superstar is",
      nba_tidy%>%
  select(player,prob_superstar,position) %>%
  filter(position=="C")%>%
  arrange(-prob_superstar)%>%
  head(1)[1,1])
```

```
## [1] "The name of the Center with the highest probability of becoming a superstar is Joel Embiid"
```

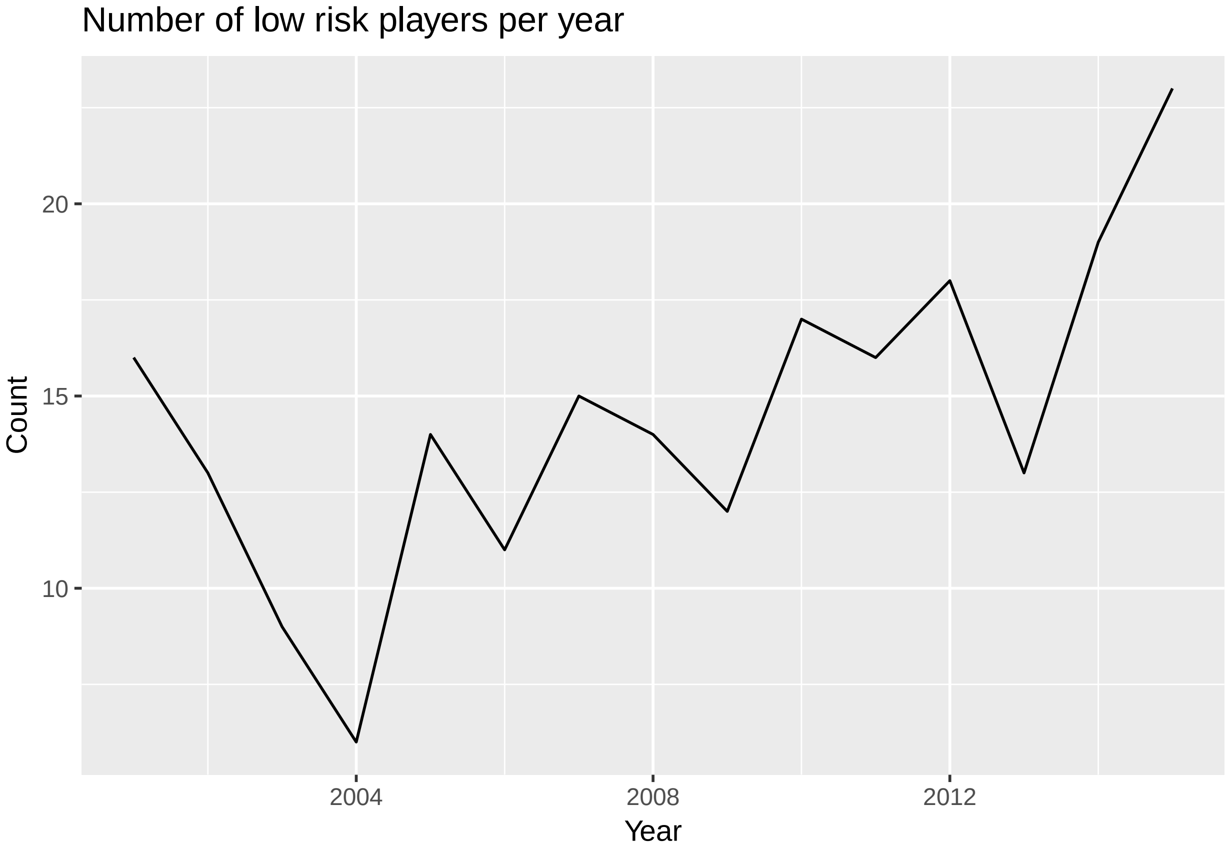
Ques 3

```
##Select postions and probability of being superstar from the tidied dataset
##group by postion and take the average of superstar probabilities for that position
##create the bar graph
nba_tidy%>%
  select(position,prob_superstar) %>%
  group_by(position)%>%
  summarize(avg_prob=mean(prob_superstar))%>%
  ggplot(aes(x=reorder(position,-avg_prob),y=avg_prob))+
  geom_bar(stat="identity")+
  labs(x="Position",y="Probability",title="Probability of becoming a superstar by position")
```



Ques 4

```
##Add a variable low_risk using mutate which is 1 for low risk and 0 otherwise
##Count the number of low risk players through the low_risk variable for each draft year
##Plot the line graph
nba_tidy%>%
  mutate(low_risk=if_else(prob_bust<0.40,1,0)) %>%
  group_by(draft_year)%>%
  summarize(count=sum(low_risk))%>%
  ggplot(aes(x=draft_year,y=count))+
  geom_line()+
  labs(x="Year",y="Count",title="Number of low risk players per year")
```



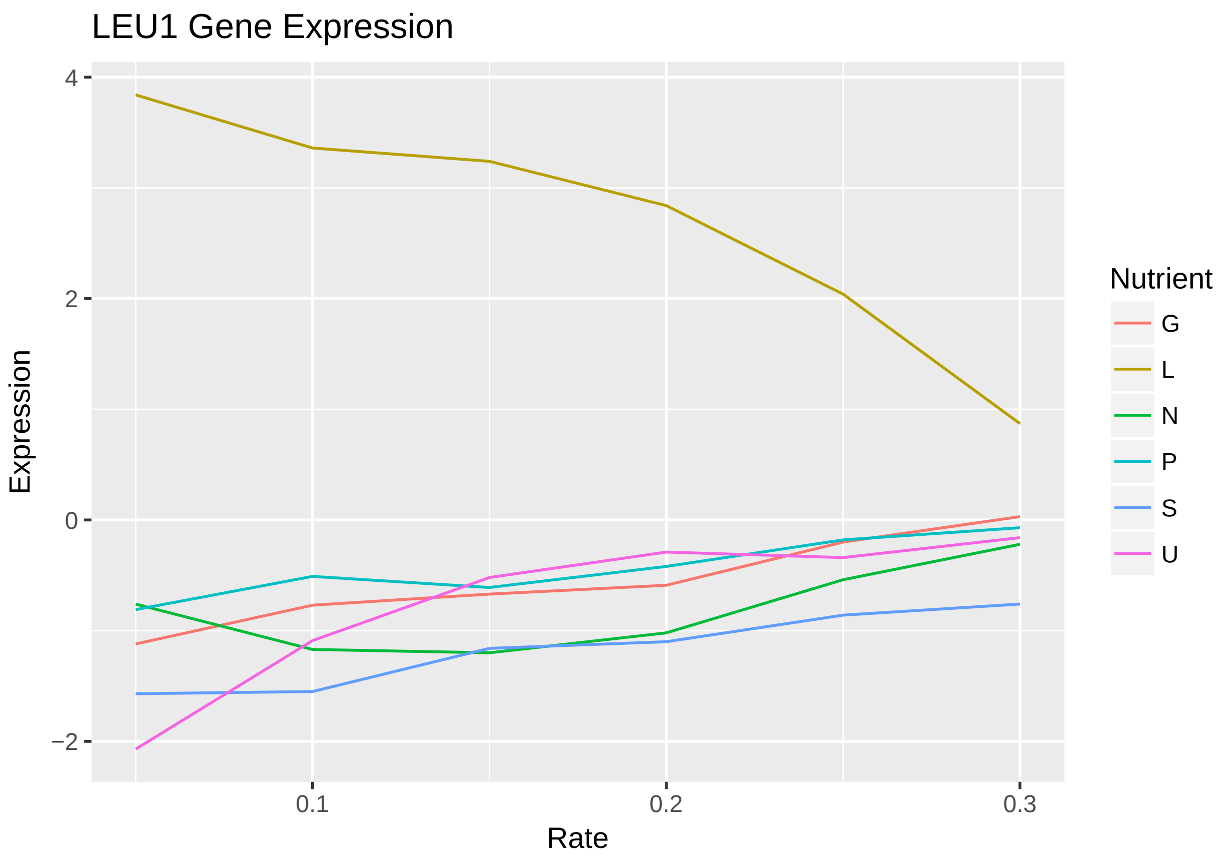
Ques 5

```
gene_data=read.csv("gene_expression.csv")
##The data is not Tidy as variable "Name" contains information
##about 2 variables i.e. gene name and the biological process
## Further, the variables G0.05...U0.3 contains information
##about 2 variables i.e. nutrient and a given growth rate
## Tidy the data by separating the Name variable into Name and bioprocess
##Gather the various nutrient & growth columns into
##nutrient_growth and then separate the new column into nutrient and growth
```

```
gene_data %>%
  separate(col=Name,into = c("Name","bio_process"),sep=":")%>%
  gather(G0.05:U0.3,key=nutrient_growth,value=expression)%>%
  separate(col=nutrient_growth,into=c("nutrient","growth"),sep=1) ->tidy_gene_data
##Remove the white spaces at end of "Name""
tidy_gene_data$Name = trimws(tidy_gene_data$Name,which="right")
##change the class of various columns
tidy_gene_data$bio_process=as.factor(tidy_gene_data$bio_process)
tidy_gene_data$nutrient=as.factor(tidy_gene_data$nutrient)
tidy_gene_data$growth=as.numeric(tidy_gene_data$growth)
##remove NAs from the expression column
tidy_gene_data%>%
  filter(!(expression=="NA")) ->tidy_gene_data
```

Ques 6

```
##Filter for LEU1 Gene, summarize the average value of
##expression for various growth rate and nutrients
##plot the line graphs
tidy_gene_data%>%
  filter(Name=="LEU1")%>%
  group_by(growth,nutrient)%>%
  summarize(avg_value=mean(expression))%>%
  ggplot(aes(x=growth,y=avg_value,group=nutrient))+
  geom_line(aes(color=nutrient))+
  labs(x="Rate",y="Expression",title="LEU1 Gene Expression",color="Nutrient")
```


LEU1 Gene Expression
4
2
0
−2
Expression
0.1
0.2
0.3
Rate
Nutrient
G
L
N
P
S
U