Homework 03

Hafsa Dawood

October 16, 2017

Case 01

1. The given dataset is not tidy. Specify the reason, and make it tidy

The given data set is not tidy because the probabilities of the player being a superstar, starter, role_player and bust are placed in one single column

```r
# Load the necessary libaries
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
```

```r
library(ggplot2)
library(tidyr)
library(stringr)


# Read the .csv file
nba_draft = read.csv("nba_draft.csv")

# Separate probabilities column into different categories
nba_draft = nba_draft %>%
             separate(probabilities, c ("superstar", "starter", "role_player", "bust"), sep = ",")


# Convert the probabilities into numeric
nba_draft$superstar = as.numeric(nba_draft$superstar)
nba_draft$starter = as.numeric(nba_draft$starter)
nba_draft$role_player = as.numeric(nba_draft$role_player)
nba_draft$bust = as.numeric(nba_draft$bust)
```

2. What was the name of the Center (position of "C") with the highest probability of becoming a superstar?

```r
# Filter by position "C", arrange in descending order of prob. of superstar and choose first row
nba_draft %>%
  filter(position == "C") %>%
  arrange(desc(superstar)) %>%
  select(player, position, superstar) %>%
  head(1)
```

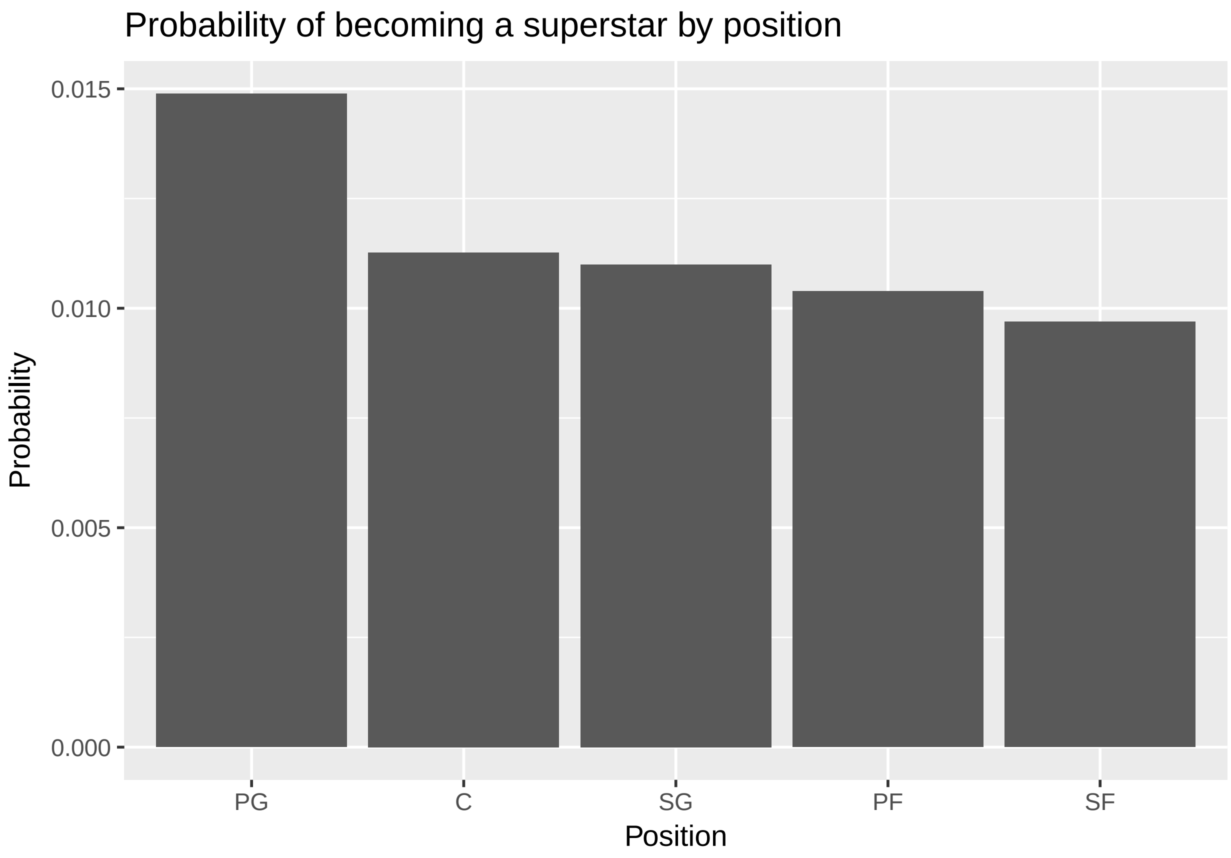
```
##        player position superstar
## 1 Joel Embiid        C 0.1478667
```

The player “Joel Embiid” has the highest probability (0.1478667) of becoming a superstar

3. Create an EXACT copy of the following graph of the average probability of becoming a superstar by

position

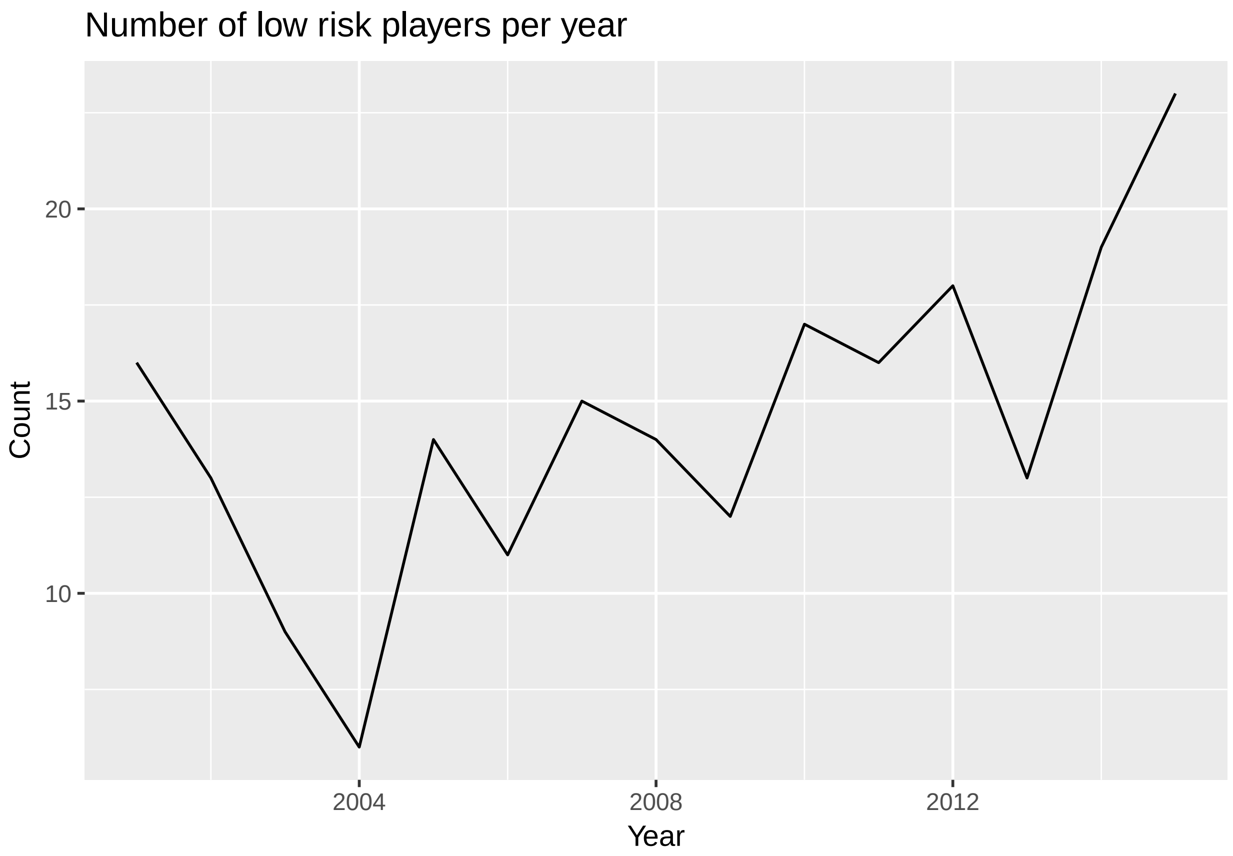
```
# Group by position and obtain mean of prob. of superstar
# In ggplot let x = position, y = avg and reorder postion based on the decreasing order of avg
# give the appropriate title and lables to the plot

nba_draft %>%
  group_by(position) %>%
  summarise(avg = mean(superstar)) %>%
  ggplot(aes(x = reorder(position,desc(avg)), y = avg)) +
  geom_bar(stat = "identity") +
  ggtitle("Probability of becoming a superstar by position") +
  xlab("Position") +
  ylab("Probability")
```



4. Create an EXACT copy of the following graph of the number of "low risk" players drafted each year.

Define "low risk" as players with less than .4 probability of being a bust.

```
# Filter for bust less than 0.4, group by the draft_year and obtain count using summarize
# Plot a line graph with x = draft_year and y = count of players
# Give the appropriate title and labels for the plot

nba_draft %>%
  filter(bust < 0.4) %>%
  group_by(draft_year) %>%
  summarize(count = n()) %>%
  ggplot(aes(x = draft_year, y=count))+
  geom_line() +
  ggtitle("Number of low risk players per year") +
  xlab("Year") +
  ylab("Count")
```



Case 02

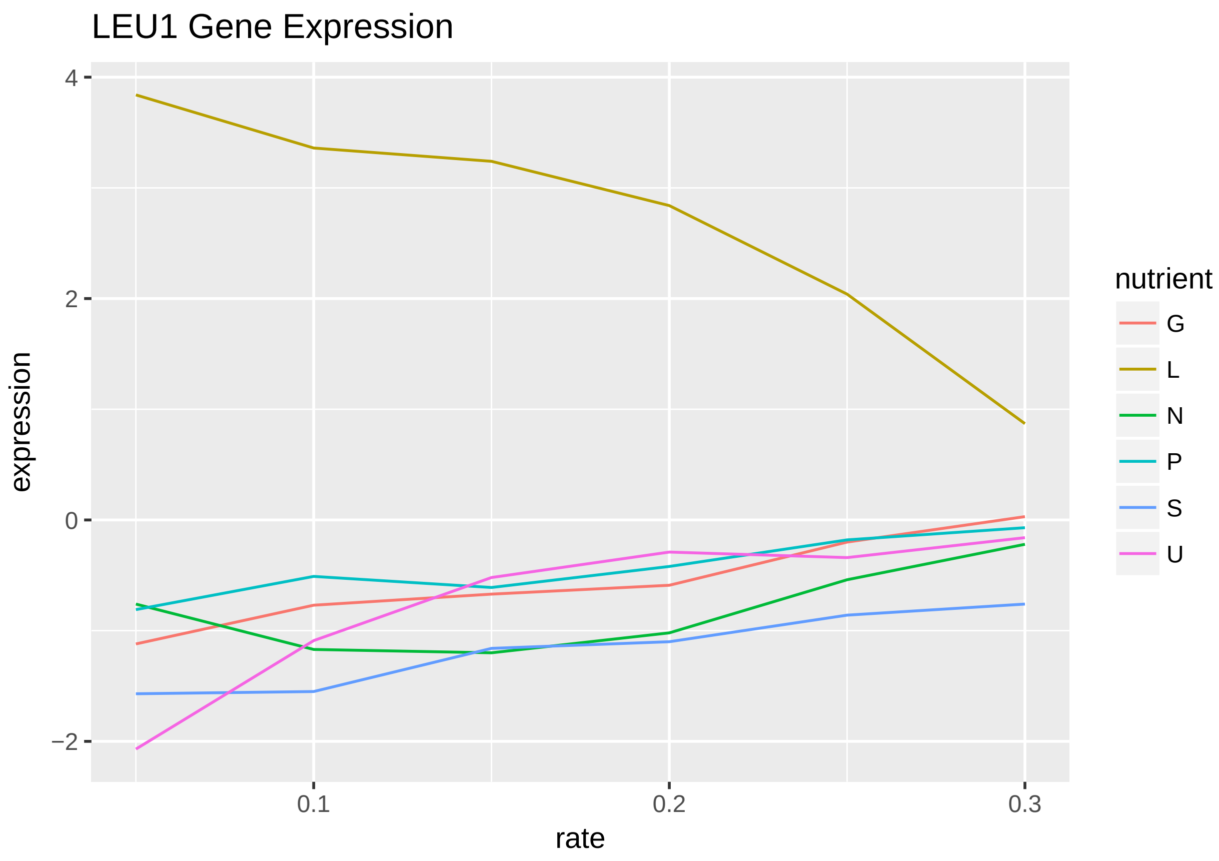
5. Is the data tidy? If not, specify why, and tidy it up.

The data is not tidy because: (i) the gene name and the biological process are present in the same column (ii) given the structure of the variable names G0.05 to U0.3 these are likely values and not variables.

```r
# Read the .csv file

gene_expression = read.csv("gene_expression.csv")

# Gather all the columns from G0.05 to U0.3
# Separate the column "Name" into gene name ("name") and biological process ("bp")
# Separate the gathered columns into "nutrient" and "growth_rate"

gene_expression = gene_expression %>%
                    gather(G0.05:U0.3, key ="key", value = "expression", na.rm = T) %>%
                    separate(Name, c("name", "bp"), sep = ":") %>%
                    separate(key, c ("nutrient", "growth_rate"), sep = 1)

# Remove leading and trailing spaces from "name" and "bp"
gene_expression$name = str_trim(gene_expression$name)
gene_expression$bp = str_trim(gene_expression$bp)

# Convert the growth rate to numeric
gene_expression$growth_rate = as.numeric(gene_expression$growth_rate)
```

6. Create the an copy of the following graph:

```r
# Filter for gene name = "LEU1"
# group by nutrient and growth rate and find the mean
# Plot a line graph, with x = growth rate and y = avg of gene expression
# Different categories of nutrients are represented by different colors
# Give the appropriate title and lables to the graph

gene_expression %>%
  filter(name == "LEU1") %>%
  group_by(nutrient,growth_rate) %>%
  summarize(avg = mean(expression)) %>%
  ggplot(aes(x = growth_rate, y = avg, color = nutrient))+
  geom_line() +
  ggtitle("LEU1 Gene Expression") +
  xlab("rate") +
  ylab("expression")
```


LEU1 Gene Expression
expression
4
2
0
−2
0.1
0.2
0.3
rate
nutrient
G
L
N
P
S
U