Homework 03

DSO 545: Statistical Computing and Data Visualization

Fall 2017

Due Date: Sunday Oct 22, 2017 (11:59 pm)

Instructions

- Please use R Markdown Documents to answer the following questions.
- Use R, dplyr, ggplot2, tidyr from the tidyverse to answer all questions. Write R code for each question.
- Make sure to include the R code you used. You won't receive any credit if you don't show the R code chunks.
- Submit your R Markdown and (pdf or word) documentation to blackboard
- I won't tolerate any kind of cheating or late submissions. However I highly encourage to disucss the assignment with each other, but make sure that everyone has a different write up.
- Good luck!

Case 01

The data set `nba_draft.csv` contains information about the NBA drafts from 2001 to 2015.

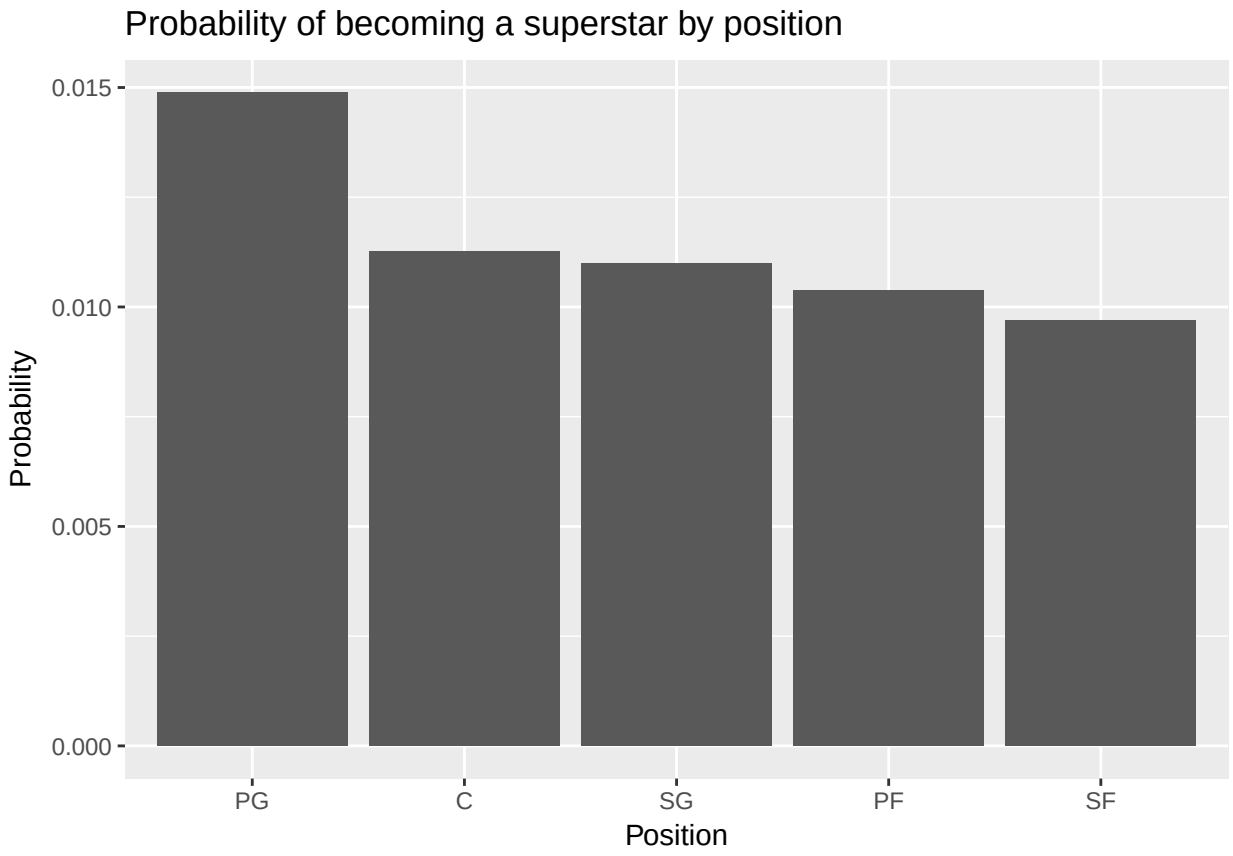
Variables

Variable	Description
player	Player's name
position	Player's position
draft_year	Year the player was drafted
probabilities	Probability the player is a `superstar`, `starter`, `role_player`, or `bust`, respectively

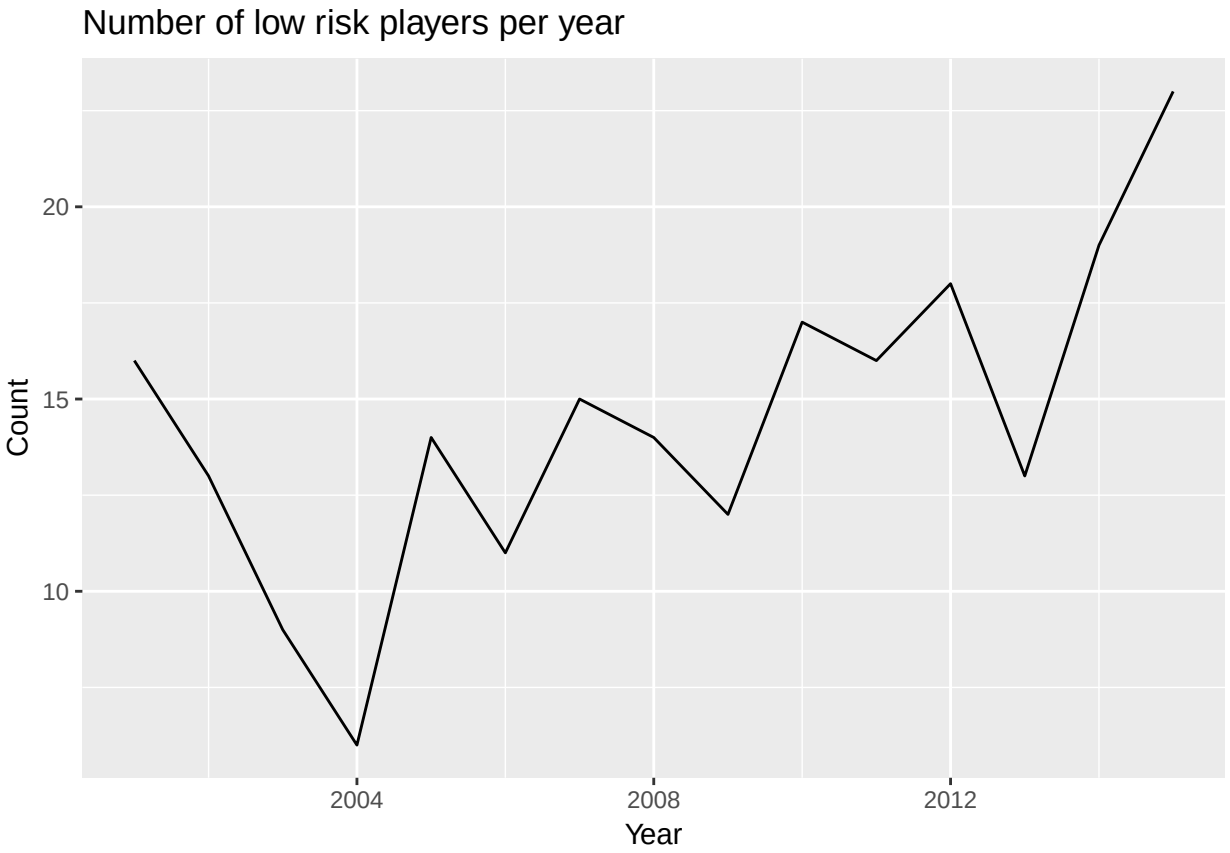
(1) The given dataset is not tidy. Specify the reason, and make it tidy.

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

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


Probability of becoming a superstar by position
0.015
0.010
0.005
0.000
Probability
PG
C
SG
PF
SF
Position

(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.



Case 02

According to wikipedia, Gene expression is the process by which the heritable information in a gene, the sequence of DNA base pairs, is made into a functional gene product, such as protien or RNA. The basic idea is that the DNA is transcribed into RNA, which is then translated into protiens. Protiens make many of the structures and all the enzymes in a cell or organism.

In this case, we are interested to learn how both the growth rate and the limiting nutrient affect the gene's expression.

Variables

Variable	Description
Name	This varaible consists of gene name ("name") and biological process ("BP") separated with a colon
G0.05,. . . , U0.3	These variables show the gene "expression" at a given "nutrient" (glucose (G), ammonium (N), sulfate (S), phosphate (P), uracil (U) or leucine (L)), and a given "Growth rate" (A number, ranging from .05 to .3. .05 means slow growth (the yeast were being starved hard of that nutrient) while .3 means fast growth)

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

(6) Create the an copy of the following graph: