Assignment: Messy Data

2019-11-03

*Group Work By: Japneet Kohli, Kevin Omores, and Prateek Chitpur*

**Data Cleaning and Preparation**

Both the datasets were cleaned manually since they were very small. The CSV files were opened and combined in Excel for ease of use, where all the data elements were visually inspected to make sure they conformed with the intended meaning of each row and column in the dataset.

There were several issues identified with raw data, both in terms of the variety in which the data was gathered as well as the meaning behind the kind of data that was captured in a particular variable. For example, “gradudate” column had responses such as the year of graduation or the semester of graduation. For clarity, this column was divided into two columns, one to capture the year and the other to hold the month of graduation information.

Similarly, there were no rules as to how the data for “gender”, “height”, “program”, or any other column were to be collected.

We tried to create a conforming scheme for all the columns by defining a data dictionary for this dataset, which identified the names of the columns, the type of data it contained, and any set standards for how the data is input. This dictionary is outlined below:

**section**: alphanumeric, S001/S003

**age**: number, in years

**height**: number, in centimeter

**gender**: alphanumeric, M/F

**Country**: alphanumeric, First letter capitalized, standardized acronyms like USA

Split **Program** into two columns:

**typeofprogram**: alphanumeric, Standardize program names

**major**: alphanumeric, major area of study

**concentration**: alphanumeric, concentration within major area of study

Split **gradudate** into two columns:

**gradudatemonth**: number, graduation month

**gradudateyear**: number, graduation year

**working**: alphanumeric, Y/N

**Notes**:

If data is missing, leave field blank.

Data that was input incorrectly and could not be assigned to any existing categories was removed. For example, a response of "Y" for gradudate was removed as bad data.

The above data dictionary summarizes the standards we assigned for preparing the data for analysis. These standards are also what we recommend to be handed out to the students in the future to make sure that the data is recorded in a more clear and concise manner. Usually, the business analysts and the data owners sketch out these schema requirements at the beginning of a data project.

**Data Summary and Visualization**

Below is a summary of the various analysis we performed on the dataset. It includes summary statistics for all the columns, as well as visualizations that give a glimpse of the overall data and also allow comparisons to be made between the two class sections.

setwd("E:/GMU/AIT-580/Classwork")  
  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts -------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

messydata <- read.csv("F19classdata001.csv", sep = ",")  
  
#Tukeys five number summary for each variable  
summary(messydata)

## section age height gender country   
## S001:26 Min. :20.00 Min. :152.0 : 4 India :25   
## S003:24 1st Qu.:23.00 1st Qu.:163.0 F:21 China : 8   
## Median :25.00 Median :170.0 M:25 USA : 7   
## Mean :26.82 Mean :169.5 : 2   
## 3rd Qu.:28.75 3rd Qu.:176.5 Taiwan : 2   
## Max. :56.00 Max. :188.0 Egypt : 1   
## (Other): 5   
## typeofprogram major   
## Graduate Certificate: 4 : 5   
## Graduate Pathway : 6 AIT : 1   
## Master's of Science :39 Data Analytics:44   
## Non Degree : 1   
##   
##   
##   
## concentration gradudateyear gradudatemonth working  
## :48 Min. :2016 Min. : 5.000 : 1   
## Data Mining : 1 1st Qu.:2020 1st Qu.: 5.000 N:32   
## Financial Engineering: 1 Median :2021 Median : 5.000 Y:17   
## Mean :2020 Mean : 8.316   
## 3rd Qu.:2021 3rd Qu.:12.000   
## Max. :2022 Max. :12.000   
## NA's :7 NA's :31

#With the summary of section, we will get total count of students in each section.  
summary(messydata$section)

## S001 S003   
## 26 24

#Summary gives mean, minimum and maximum age of students of both the sections.  
summary(messydata$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 23.00 25.00 26.82 28.75 56.00

#Summary gives mean, minimum and maximum height of students of both the sections  
summary(messydata$height)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 152.0 163.0 170.0 169.5 176.5 188.0

#Summary gives count of male and female students of both the sections  
summary(messydata$gender)

## F M   
## 4 21 25

#Summary gives count of students from different countries of both the sections  
summary(messydata$country)

## China Egypt India Iran Ivory Coast   
## 2 8 1 25 1 1   
## Kazakhstan Palestine Taiwan Thailand USA   
## 1 1 2 1 7

#Summary gives count of types of programs in which students enrolled of both the sections  
summary(messydata$typeofprogram)

## Graduate Certificate Graduate Pathway Master's of Science   
## 4 6 39   
## Non Degree   
## 1

#Summary shows 1 student belonged to different major  
summary(messydata$major)

## AIT Data Analytics   
## 5 1 44

#Summary gives count of students who chose concentration or not  
summary(messydata$concentration)

## Data Mining Financial Engineering   
## 48 1 1

#Summary gives mean value of graduation year  
summary(messydata$gradudateyear)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 2016 2020 2021 2020 2021 2022 7

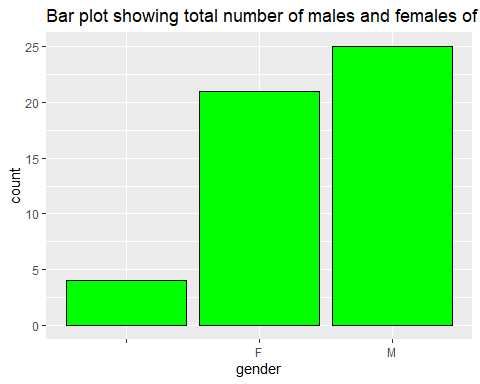
#Summary gives count of students who are working or not working of both the sections  
summary(messydata$working)

## N Y   
## 1 32 17

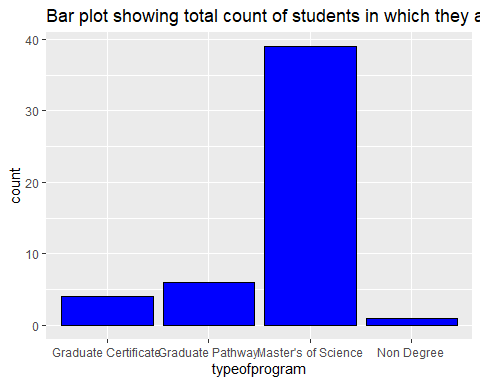
#graduatedmonth contains 31 NAs, which is more than 25%. So, the variable is neglected for visualizations.  
sum(is.na(messydata$gradudatemonth))

## [1] 31

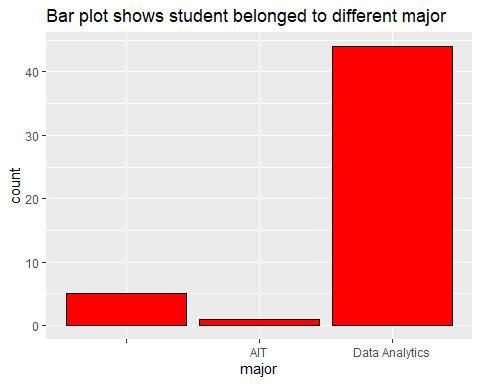
#Visualizations of analysis using barplot.  
#Note: Results contain count of missing values or unknown values  
  
ggplot(messydata, aes(x = gender)) + geom\_bar(fill = "green", color = "black") +  
 labs(title = "Bar plot showing total number of males and females of both sections")



ggplot(messydata, aes(x = typeofprogram)) + geom\_bar(fill = "blue", color = "black")+  
 labs(title = "Bar plot showing total count of students in which they are enrolled to program")

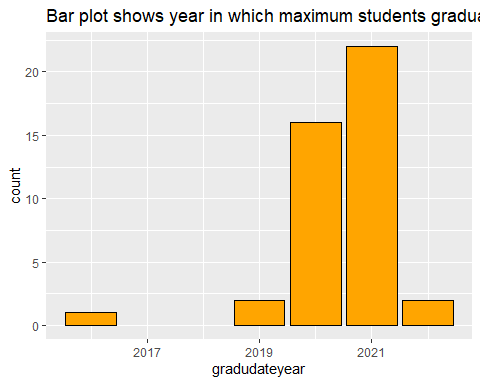


ggplot(messydata, aes(x = major)) + geom\_bar(fill = "red", color = "black")+  
 labs(title = "Bar plot shows student belonged to different major")

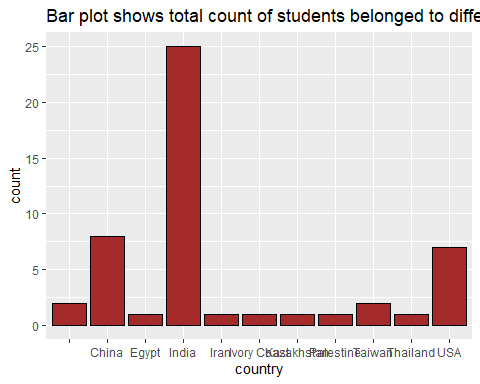


ggplot(messydata, aes(x = gradudateyear)) + geom\_bar(fill = "orange", color = "black")+  
 labs(title = "Bar plot shows year in which maximum students graduate")

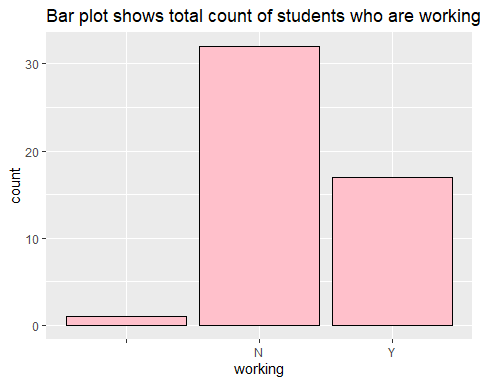
## Warning: Removed 7 rows containing non-finite values (stat\_count).



ggplot(messydata, aes(x = country)) + geom\_bar(fill = "brown", color = "black")+  
 labs(title = "Bar plot shows total count of students belonged to different country")



ggplot(messydata, aes(x = working)) + geom\_bar(fill = "pink", color = "black")+  
 labs(title = "Bar plot shows total count of students who are working(Y) or not working(N)")



#messydatasection1 and messydatasection2 are two subsets of main data frame messydata  
  
messydatasection1 <- messydata[messydata$section == 'S001',]  
messydatasection1

## section age height gender country typeofprogram  
## 1 S001 48 177 USA Graduate Certificate  
## 2 S001 28 165 F Master's of Science  
## 3 S001 34 170 M Egypt Master's of Science  
## 4 S001 24 168 M India Master's of Science  
## 5 S001 23 173 M India Master's of Science  
## 6 S001 25 178 M USA Master's of Science  
## 7 S001 24 163 F China Master's of Science  
## 8 S001 27 168 M Non Degree  
## 9 S001 30 178 M Ivory Coast Master's of Science  
## 10 S001 34 160 F Iran Master's of Science  
## 11 S001 23 178 India Master's of Science  
## 12 S001 27 183 M Thailand Graduate Pathway  
## 13 S001 22 168 F China Master's of Science  
## 14 S001 56 188 M USA Graduate Certificate  
## 15 S001 23 170 F India Master's of Science  
## 16 S001 29 157 F India Master's of Science  
## 17 S001 22 180 M India Master's of Science  
## 18 S001 22 175 M India Master's of Science  
## 19 S001 23 180 M India Master's of Science  
## 20 S001 23 152 F India Master's of Science  
## 21 S001 25 170 M Kazakhstan Master's of Science  
## 22 S001 23 170 India Master's of Science  
## 23 S001 26 152 M India Master's of Science  
## 24 S001 29 153 F India Master's of Science  
## 25 S001 23 165 M India Master's of Science  
## 26 S001 24 170 F China Master's of Science  
## major concentration gradudateyear gradudatemonth working  
## 1 Data Analytics 2020 5 Y  
## 2 Data Analytics 2021 NA N  
## 3 Data Analytics 2021 NA Y  
## 4 AIT 2019 12 N  
## 5 Data Analytics 2021 5 N  
## 6 Data Analytics 2022 NA Y  
## 7 Data Analytics 2020 12 N  
## 8 NA NA Y  
## 9 Data Analytics 2020 12 Y  
## 10 Data Analytics 2021 NA Y  
## 11 Data Analytics 2021 NA N  
## 12 NA NA N  
## 13 Data Analytics 2020 12 N  
## 14 Data Analytics 2019 NA Y  
## 15 Data Analytics Data Mining 2020 NA Y  
## 16 Data Analytics 2020 NA N  
## 17 Data Analytics 2020 12 N  
## 18 Data Analytics 2021 5 N  
## 19 Data Analytics 2020 12 N  
## 20 Data Analytics 2021 NA N  
## 21 Data Analytics 2021 NA N  
## 22 Data Analytics 2021 NA N  
## 23 Data Analytics 2020 5 Y  
## 24 Data Analytics 2021 5 N  
## 25 Data Analytics 2021 5 N  
## 26 Data Analytics 2020 12 N

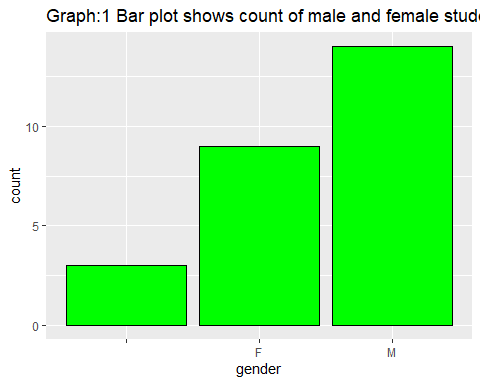
messydatasection2 <- messydata[messydata$section == 'S003',]  
messydatasection2

## section age height gender country typeofprogram major  
## 27 S003 24 170 F China Master's of Science Data Analytics  
## 28 S003 33 162 M India Master's of Science Data Analytics  
## 29 S003 22 175 M USA Master's of Science Data Analytics  
## 30 S003 20 173 F India Master's of Science Data Analytics  
## 31 S003 23 157 F India Master's of Science Data Analytics  
## 32 S003 34 173 M China Graduate Pathway   
## 33 S003 31 160 F India Master's of Science Data Analytics  
## 34 S003 31 167 F China Graduate Certificate Data Analytics  
## 35 S003 36 180 M USA Graduate Certificate Data Analytics  
## 36 S003 27 165 F India Graduate Pathway Data Analytics  
## 37 S003 27 172 M India Master's of Science Data Analytics  
## 38 S003 26 185 M USA Master's of Science Data Analytics  
## 39 S003 23 160 F India Master's of Science Data Analytics  
## 40 S003 26 172 M Taiwan Master's of Science Data Analytics  
## 41 S003 23 160 F India Master's of Science Data Analytics  
## 42 S003 23 160 Taiwan Graduate Pathway Data Analytics  
## 43 S003 25 157 F India Master's of Science Data Analytics  
## 44 S003 27 163 F India Master's of Science Data Analytics  
## 45 S003 30 165 F Palestine Master's of Science Data Analytics  
## 46 S003 27 183 M USA Master's of Science Data Analytics  
## 47 S003 21 188 M India Master's of Science Data Analytics  
## 48 S003 21 163 F India Master's of Science Data Analytics  
## 49 S003 22 178 M China Graduate Pathway   
## 50 S003 22 175 M China Graduate Pathway   
## concentration gradudateyear gradudatemonth working  
## 27 2020 12 N  
## 28 2021 5 N  
## 29 2020 NA Y  
## 30 2021 5 N  
## 31 2021 5 N  
## 32 NA NA Y  
## 33 2021 NA N  
## 34 2020 NA Y  
## 35 2020 NA Y  
## 36 NA NA   
## 37 2021 5 Y  
## 38 Financial Engineering 2021 NA Y  
## 39 2021 NA N  
## 40 2016 NA N  
## 41 2021 NA N  
## 42 NA NA N  
## 43 NA NA Y  
## 44 2021 NA N  
## 45 2020 12 Y  
## 46 2020 NA N  
## 47 2021 NA N  
## 48 2021 NA N  
## 49 NA NA N  
## 50 2022 NA N

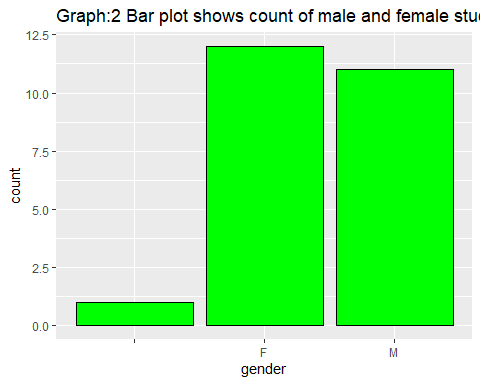
str(messydatasection1)

## 'data.frame': 26 obs. of 11 variables:  
## $ section : Factor w/ 2 levels "S001","S003": 1 1 1 1 1 1 1 1 1 1 ...  
## $ age : int 48 28 34 24 23 25 24 27 30 34 ...  
## $ height : int 177 165 170 168 173 178 163 168 178 160 ...  
## $ gender : Factor w/ 3 levels "","F","M": 1 2 3 3 3 3 2 3 3 2 ...  
## $ country : Factor w/ 11 levels "","China","Egypt",..: 11 1 3 4 4 11 2 1 6 5 ...  
## $ typeofprogram : Factor w/ 4 levels "Graduate Certificate",..: 1 3 3 3 3 3 3 4 3 3 ...  
## $ major : Factor w/ 3 levels "","AIT","Data Analytics": 3 3 3 2 3 3 3 1 3 3 ...  
## $ concentration : Factor w/ 3 levels "","Data Mining",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ gradudateyear : int 2020 2021 2021 2019 2021 2022 2020 NA 2020 2021 ...  
## $ gradudatemonth: int 5 NA NA 12 5 NA 12 NA 12 NA ...  
## $ working : Factor w/ 3 levels "","N","Y": 3 2 3 2 2 3 2 3 3 3 ...

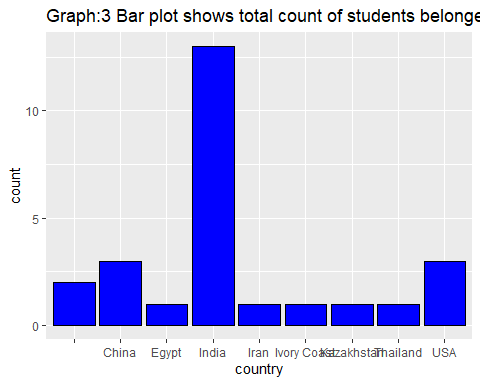
#We can compare from below graphs 1 & 2 that section 1 has maximum male students, whereas section 2 has maximum female students.  
#Also graph shows count of unknown values  
ggplot(messydatasection1, aes(x = gender)) + geom\_bar(fill = "green", color = "black")+  
 labs(title = "Graph:1 Bar plot shows count of male and female students of section1")



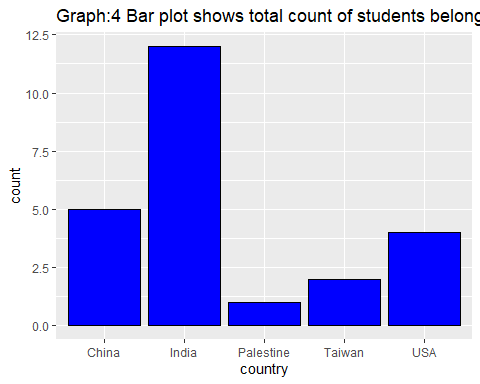
ggplot(messydatasection2, aes(x = gender)) + geom\_bar(fill = "green", color = "black")+  
 labs(title = "Graph:2 Bar plot shows count of male and female students of section2")



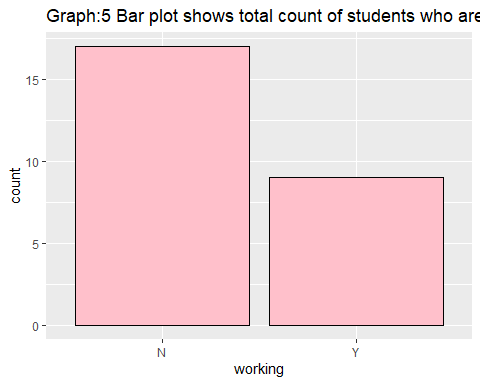
#It is observed from below both the graphs 3 & 4 that count of Indian students is maximum in both sections  
#Also graph conains count of missing values  
ggplot(messydatasection1, aes(x = country)) + geom\_bar(fill = "blue", color = "black")+  
 labs(title = "Graph:3 Bar plot shows total count of students belonged to different country in section1")



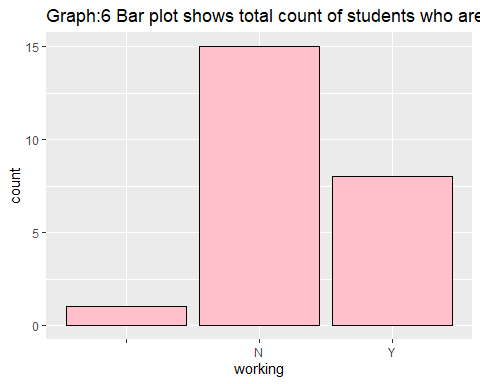
ggplot(messydatasection2, aes(x = country)) + geom\_bar(fill = "blue", color = "black")+  
 labs(title = "Graph:4 Bar plot shows total count of students belonged to different country in section2")



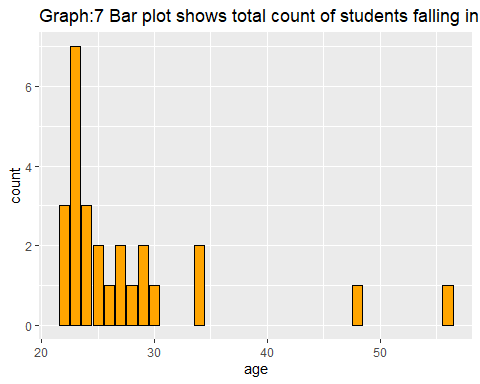
#We can observe from below graphs 5 & 6 that section1 has maximum students who are not working as compared to section2  
#Graph depicts count of unknown values as well  
ggplot(messydatasection1, aes(x = working)) + geom\_bar(fill = "pink", color = "black")+  
 labs(title = "Graph:5 Bar plot shows total count of students who are working(Y) or not working(N) in section1")



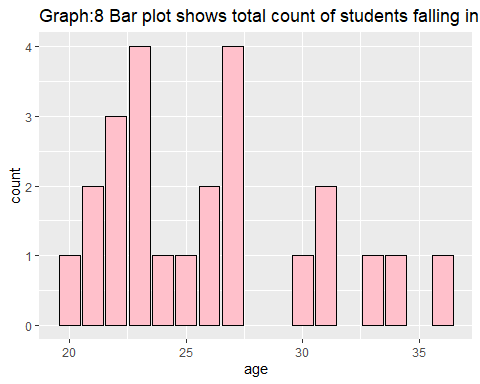
ggplot(messydatasection2, aes(x = working)) + geom\_bar(fill = "pink", color = "black")+  
 labs(title = "Graph:6 Bar plot shows total count of students who are working(Y) or not working(N) in section2")



#From below both graphs 7 & 8, we can observe that both sections has maximum students of age 23  
ggplot(messydatasection1, aes(x = age)) + geom\_bar(fill = "orange", color = "black")+  
 labs(title = "Graph:7 Bar plot shows total count of students falling in different age of section1")

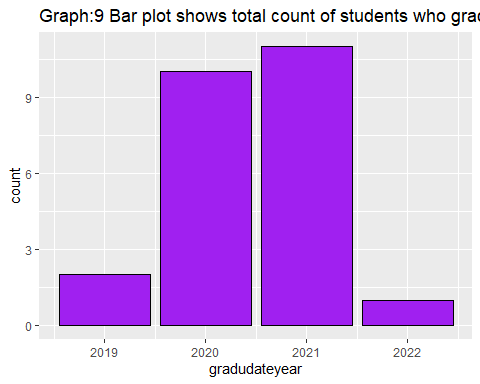


ggplot(messydatasection2, aes(x = age)) + geom\_bar(fill = "pink", color = "black")+  
 labs(title = "Graph:8 Bar plot shows total count of students falling in differen age of section2")



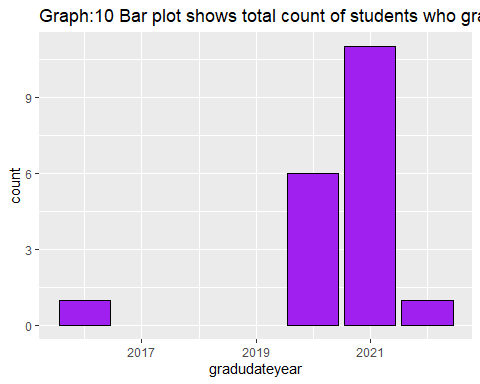
#We can observe from below both graphs 9 & 10 that maximum students graduate in the year 2021 in both the sections  
ggplot(messydatasection1, aes(x = gradudateyear)) + geom\_bar(fill = "purple", color = "black")+  
 labs(title = "Graph:9 Bar plot shows total count of students who graduate in particular year of section1")

## Warning: Removed 2 rows containing non-finite values (stat\_count).

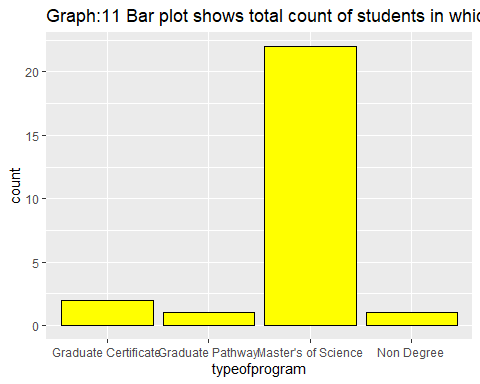


ggplot(messydatasection2, aes(x = gradudateyear)) + geom\_bar(fill = "purple", color = "black")+  
 labs(title = "Graph:10 Bar plot shows total count of students who graduate in particular year of section2")

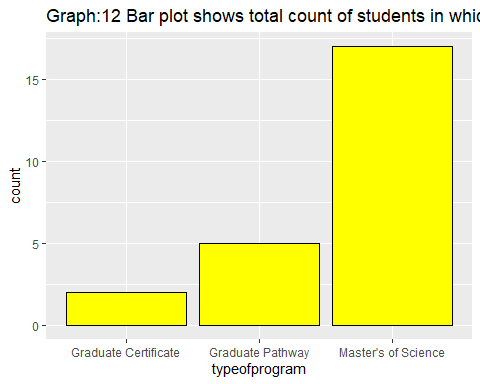
## Warning: Removed 5 rows containing non-finite values (stat\_count).



#It's observed from below both graphs 11 & 12 that section 2 has more number of pathway students than section 1   
ggplot(messydatasection1, aes(x = typeofprogram)) + geom\_bar(fill = "yellow", color = "black")+  
 labs(title = "Graph:11 Bar plot shows total count of students in which they are enrolled to program of section1")



ggplot(messydatasection2, aes(x = typeofprogram)) + geom\_bar(fill = "yellow", color = "black")+  
 labs(title = "Graph:12 Bar plot shows total count of students in which they are enrolled to program of section2")



**Comments and Suggestions**

The exercise seemed to mirror a real-life data collection and analysis project to a certain degree. We worked in a team, and a basis of how we were going to tackle this assignment was sent in the classroom itself. We divided the work based on the different areas of work we identified and collaborated remotely to complete the assignment.

We think it works better if this is made to be a mandatory group assignment, especially considering that all the other work done in this class is individual work.