Exam 1

Add name

## Introduction

Congratulations! You have just been hired to the California state education department as a data analyst. Your boss has an initiative to improve college advising for an online advising tool. To get started she needs data to understand California colleges and academic programs available to recommend colleges for different individuals.

Please show each step with comments in the R code and solution formatted correctly in a brief sentence.

## 1. The first request is to add a sample of users to analyze:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Full Name | Birthday | Male | Priority | Program of interest |
| MCPHERSON, JOHN | 2/2/2002 | 1 | Low cost college on campus | Physical Science |
| PERCY, JUDY | 3/3/1998 | 0 | Online college with high graduation rate | Communication |
| LONG, ELENA | 4/4/2003 | 0 | High earnings on campus | Architecture |
| THAYER, RHONDA | 5/5/2004 | 0 | High earnings to cost ratio on campus | Engineering |

#### 1.1 Create a data frame (df\_sample) for a sample of the 4 users on the online tool. (1 pt)

# Create data frame

#### 1.2 The Male = 1 indicate the user is a male. What is the probability of being a male based on this sample? (1 pt)

#### 1.3 Add an indicator if the user is currently in high school based on birthday between September 2001 and September 2004. (2 pts)

# Add indicator for high school

## 2. The second request is to review and clean publically available data:

load("college\_scorecard\_ca\_completion.Rda")  
load("college\_scorecard\_ca\_programs.Rda")  
load("college\_scorecard\_data\_dictionary.Rda")

#### 2.1 Based on the data\_dictionary what is the data element full name for the following variables? (1pt)

* PCIP27
* DISTANCEONLY
* MD\_EARN\_WNE\_P10
* C150\_4\_POOLED

#### 2.2 Create a new column called npt that combines npt4\_pub & npt\_priv. (1pt)

# Create new column

#### 2.3 Filter colleges (2pt)

Filter the data for (1) Predominantly bachelor’s-degree granting colleges (preddeg==3) and (2) colleges without NA values for all the following variables: (1pt)

* distance\_only
* npt from 2.2
* md\_earn\_wne\_p10
* c150\_4\_pooled

# Filter the data

How many colleges of the original data set were filtered out? What percent of colleges remain Write a brief sentence to your boss explaining the cleaning of the data. (1pt)

## 3. The third request is to select a calculate a few statistics: (5pts)

#### 3.1 Tuition Price

The median tuition price is \_\_\_\_\_\_\_\_\_\_ with the lowest cost college at \_\_\_\_\_\_\_\_\_\_ with a cost of \_\_\_\_\_\_\_\_\_\_.

#### 3.2 Graducation

The mean graduation rate is \_\_\_\_\_\_\_\_\_\_ with the highest graduation rate at \_\_\_\_\_\_\_\_\_\_ with a rate of \_\_\_\_\_\_\_\_\_\_.

#### 3.3 Future earnings per tuition Price

The highest ratio of future earnings per tuition price is at \_\_\_\_\_\_\_\_\_\_ with a ratio of \_\_\_\_\_\_\_\_\_\_.

## Quality review (2pt)

Stop here and review your report so far.

* Make sure to add your name to the report
* Does it knit correctly?
* Are there any misspellings?
* Can anything be formatted/rounded for better communication in the report?

Take 5 minutes to organize the code and rmarkdown for submission.

## Challenge! Only if you have time..

Select one of the users and determine the top 3 colleges. (2pts)  
Base the recommendation on:

* Online preference: Filter for online or on campus colleges depending on the preference indicated
* Program: Filter for colleges with at least 10% Percentage of degrees awarded in the preferred program
* Priority: Sort based on priority statistic

# Logic for recommendation