STA 135 Exam I Project, due
Tuesday, May 4^{th} in Gradescope

Read the following instructions carefully:

- **You may work in a group of two, or three.**
- **You are not allowed to discuss the questions with anyone other than the instructors (Cong or myself) and your group mate.**
- **Any outside help beyond that from the instructors is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous functions, or the instructors functions that are posted online.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- You may ask clarifying questions about code and general approach on Piazza, but do not give away any numerical answers. If you are concerned you may be giving something away, email us directly.

Table 1: Characteristics of a newly born.

length	Birthweight	headcirumference	Gestation	smoker	motherage	mnocig	mheight	mppwt	fage	fedyrs	fnocig	fheight	lowbwt	mage35
61	5.1	36	43	0	43	7	165	64	38	19	45	189	0	1

Birthweight reduced data set

This dataset contains information on new born babies and their parents. It contains mostly continuous variables (although some have only a few values e.g. number of cigarettes smoked per day). The birthweights of the babies who mothers smoked have been adjusted slightly to exaggerate the differences between mothers who smoked and didn't smoke so students can see the difference more clearly in a scatterplot with gestational age and scatter colour coded by smoking status.

Main dependent variable = Birthweight (lbs)

The data is found in the file **Birthweight-reduced-kg-R.cvs**, with the following columns:

Name	Variable	Data type
ID	Baby number	
length	Length of baby (cm)	Scale
Birthweight	Weight of baby (kg)	Scale
headcirumference	Head Circumference	Scale
Gestation	Gestation (weeks)	Scale
smoker	Mother smokes 1 = smoker 0 = non-smoker	Binary
motherage	Maternal age	Scale
mnocig	Number of cigarettes smoked per day by mother	Scale
mheight	Mothers height (cm)	Scale
mppwt	Mothers pre-pregnancy weight (kg)	Scale
fage	Father's age	Scale
fedyrs	Father's years in education	Scale
fnocig	Number of cigarettes smoked per day by father	Scale
fheight	Father's height (kg)	Scale
lowbwt	Low birth weight, 0 = No and 1 = yes	Binary
mage35	Mother over 35, 0 = No and 1 = yes	Binary

For your analysis you can ignore the last two columns. The following questions should guide your analysis and report.

- Examine appropriate plots of the data and interpret them in the context of the problem. In your interpretations, make sure to specifically indicate which individuals (babies) stand out from the rest and also discuss it i.e., what make them different from the rest.
- This part involves performing a PCA for the data using the correlation matrix. To run the PCA, you should discard the categorical variable (smoking status).
  - Discuss the positive and negative aspects of using the covariance matrix for a PCA rather than the correlation matrix.
  - Determine the number of PCs needed when using the correlation matrix.
  - Interpret the PCs chosen from the previous part. Make sure to specifically comment on whether positive or negative scores (or scores close to 0) for a PC would likely be beneficial for baby.

    Hint: To help see which of the loadings are "away from zero", one could set the cutoff argument value in summary() to something other than 0.0. For example one can use a cutoff of 0.2.
  - Examine plots of the PC scores and interpret them in the context of the problem. For example, what do you think of a particular patient (Baby).
  - Suppose a new baby is born after the PCA has been completed. The newly born has the characteristics as indicated in Table 1. Through using the previous PCA results, discuss how this particular individual/baby would compare to the other babies.
  - If the overall goal is to compare babies according with their mother's smoking status, how would you proceed ?

    My suggestion: Run PCs on these two groups (use the smoking status to split the data and run PCs on them). Then comment on their similarities and differences. You can compare their PCs, scree plots, etc ...
  - What is your general conclusion and possible future work. By future work, I mean if you have more time what possible questions would you want to investigate ?

The Report Format

You report should be writing in full sentences, and have the following sections, while being **as specific as you can** about your results. **There should not be any "copy and pasted" R code in this report. You must format the results you get from R.**

I. Introduction. State the question you are trying to answer, why it is a question of interest (why might we be interested in the answer), and what statistical technique you are going to use. Hint: use the proposed questions to formulate a question of interest.

II. Summary of your data. This should include things like plots (histograms, boxplots, star plot, parallel, etc ...) including the interpretation of the plots, and summary values such as sample means and standard deviations. This is where you should justify your choice of question of interest.

III. Analysis. Remember to write your results in full sentences where possible.

IV. Interpretation. State your conclusion, and inference that you may draw from your corresponding analysis. These should all be in terms of your problem.

V. Conclusion. Summarize briefly your findings. Here you do not have to re-iterate your numeric values, but summarize all relevant conclusions.

Details

Your report should be the following format:

i. Typed.

ii. Double-sided pages.

iii. An appendix of your R code used to produce the results. **Do not include in R code in the body of your report.**

iv. You are welcome to provide a nice/meaningful title to your project.

Notice: your project will be graded as a group effort (if you have two/three people). This means that you are responsible for your own work, and your partners work. We will not assign two different grades to one project.