The data set contains longitudinal information about 380 women who have not had a hysterectomy and have not experienced menopause before intake (recruitment). "Menopause" is ascertained when a woman has had no menstrual periods for 12 consecutive months. Henceforth this "ascertained menopause" will be the condition of interest and will be referred to simply as "menopause". All 380 women were followed over time until they experienced menopause, died, or were censored due to either the woman dropping out or the study ending. Your colleague seeks to understand which exposure(s) might influence event time (menopause time) in this population. There were 75 women who experienced menopause during the follow-up time.

The data from this study are in a flat file called menopause.DAT. There are 6 columns in the file corresponding to the following variables:
id = Id number of patient
intake_age = Patient's age (in years) when the patient was recruited into the study.

menopause_age = Patient's menopause age (in years) or censoring age (in years).

menopause = 1 if patient was observed to experience menopause;
0 if patient was censored at menopause_age.

race = 0 if patient is White non-Hispanic;
1 if patient is Black, non-Hispanic;
2 if patient is Other Ethnicity.

education = 0 if patient had Post-graduate;
1 if patient had College Graduate;
2 if patient had Some College;
3 if patient had High School Education (or less).

(A) Let's define the *menopause time* as the duration of time in the study at which the patient experienced menopause, i.e. *menopause time = menopause_age - intake_age*. We want to analyze *menopause times*.

I. Let's first study *menopause time* in the entire sample.
(Ia) Assuming that the distribution of *menopause times* for these subjects is approximately exponential, estimate the median *menopause time* for all subjects disregarding covariates. Interpret. Generate an approximate (i.e., large sample) 95% confidence interval for this parameter.
(Ib) Compute a nonparametric estimate of the survival function for *menopause time* in these data via the method of Kaplan and Meier. Provide a table of the estimates and a graph.

What is the estimated median time to menopause according to the Kaplan Meier curve? Explain the difference between your estimate based on the exponential distribution and your result here.

II. To answer the questions below, use regression techniques to find a best fitted proportional hazards model as a function of *race*, *education*, and *intake_age*. Interpret your result based on your final model and check the proportional hazard assumption.

(B) Your colleague is more interested in analyzing *menopause_age*, not *menopause_age - intake age*. It is reasonable to assume that *menopause_age* and *intake_age* are quasi-independent. Please note *menopause_age* is always no less than *intake_age* for every patient due to the sampling design. Two random variables X (*intake_age*) and T (*menopause_age*) with $P(T \geq X)$ are said to be quasi-independent iff the joint pdf $f(x,t)$ of (X,T) can be written as $f(x,t) = C\,f(t)g(x)$ for $t>x$ and $=0$ otherwise, where f and g are two pdfs and C is a constant that makes $f(x,y)$ a genuine bivariate pdf.

III. Compute a nonparametric estimate for the survival function $S(t) = \int_0^t f(x)dx$ of *menopause_age* for these data. Provide a table of the estimates and a graph. What is the median survival time according to your estimate? Explain the difference between your estimate based on the exponential distribution and your result here.

IV. Now let's look at how *race* relates to *menopause_age*.

Test whether the survival distributions for the three *race* groups are equivalent or not (you should consider that *intake_age* is always less than *menopause_age*). What do you conclude?

V. Your colleague asks whether *race* provides additional information about *menopause_age* beyond that provided by *education*. To answer the questions below, use regression techniques to model *menopause_age* as a function of *race* and *education* (you should consider that *intake_age* is always less than *menopause_age*).

(Va) Is *race* a significant predictor of *menopause_age* after adjusting for *education*?
(Vb) Provide point and 95% confidence interval estimates for the relative risk of *menopause_age* for a Black Patient with an Other Ethnicity patient controlling for *education*. Interpret.
(Vc) Based on the regression model for *menopause_age* as a function of *race* and *education*, produce an estimate of the baseline survival function for White non-Hispanic patients with Post-graduate education.
(Vd) Check your model assumptions. Specifically, if you used the proportional hazards model, please check the proportional hazards assumption.