Background: The data file WageRace.txt contains observations from the 1988 March U.S. Current Population Survey on weekly wages in 1987 (in 1992 dollars) for a sample of 25,632 males between the age of 18 and 70 who worked full-time, along with their years of education, years of experience, race (1=black, 0=not black), whether they worked in a standard metropolitan statistical area (1=in or near a city, 0=otherwise), and the region of the U.S. where they worked (MW, NE, S, W). (Source: Ramsey and Shafer, 2002)

(a) The primary research question is whether blacks have a lower salary, on average, compared to nonblacks for the population of males at this time. State the null and alternative hypothesis in terms of population means,

\[ H_0: \mu_{\text{black}} - \mu_{\text{nonblack}} = 0 \]
\[ H_a: \mu_{\text{black}} - \mu_{\text{nonblack}} > 0 \]

(b) Summarize what you learn from the initial graphical exploration of these data.

Both salary distributions are strongly skewed to the right. Hard to compare because not on the same scale.
(c) What about the following boxplots?

(d) The difference in the sample means is about $200/week. Do you find this to be a substantial difference? How large would the difference need to be to convince you that this issue is worth investigating further? [Hint: I want you to focus on the context here.] Is there a different comparison you would rather make? Explain if so.

(e) Would you recommend a two-sample t test to compare the two sample means? Explain.

(f) Suppose the two-sample t test comes out statistically significant. Write a one-sentence summary of the conclusion you would draw.

(g) Suppose the two-sample t test is not statistically significant. Does that prove that the population means are equal? Why or why not?
(h) Suggest some reasons for the large standard deviations in these data. In other words, what might be some other “sources of variation” in these salaries?

(i) Would increasing the sample sizes necessarily lower the sample standard deviations?

(j) Suggest an alternative analysis procedure you could apply rather than a two-sample t-test.

(k) Suppose we only wanted to compare the weekly wages of males with the same years of education. Suggest an advantage to doing so.

(l) Is years of education a potential “source of variability” in the weekly wages? In other words, do you expect there to be an association between years of education and weekly wages? Describe how you expect these two variables to be related. Is years of education a potential confounding variable in this study? Explain what would need to be true for it to be confounding.

(m) Suggest a way to carry out a two-sample t-test to compare the weekly wages of males with similar levels of education. Would you recommend this analysis? Explain.
Similarly we could consider the *years of experience*

(n) Do you notice any difficulties with using this variable? What would you suggest exploring in the dataset first? [Hint: I'm looking for a subtle problem with the data that these graphs reveal.]

(o) There were four geographic regions. We could do a two-sample comparison between blacks and non-blacks within each region. What is the main advantage and the main disadvantage to carrying out these four different tests?

(p) Suppose we wanted to know whether the “wage discrepancy” between blacks and non-blacks differed among the four regions. How would you suggest exploring this research question?
Summary: By the end of this course, you will have learned how to compare the weekly wages of blacks and non-blacks after adjusting for other quantitative and categorical variables like education, experience, and region. You will be able to estimate the “effect” of race when comparing individuals with similar characteristics; though still being cautious to not draw a cause-and-effect conclusion from such observational data. You will learn how to deal with skewed distributions and more than two groups. You will learn how to adjust for multiple comparisons to control the overall Type I Error rate. And you will continue to use graphical and numerical summaries to explore your data and investigate possible anomalies (like negative years of experience) and the impacts they may have on the analysis, and to distinguish between statistical and practice significance. You will also learn about the fundamental concept of interaction between variables. A common theme to these new tools will be identifying, describing, accounting for, and possibly explaining different sources of variability in the data.