This homework is due on Friday, October 25^{th} at 11.59pm. Late submissions are not accepted.

Submission Guidelines

- You must make two submissions:
  - Your complete homework as a SINGLE PDF file by the stated deadline to the gradescope. **Include your code and output of the code as texts in the PDF.**
  - Your codes to a separate submission: a single notebook file including codes for questions 6 through 12.
- For your PDF submissions:
  - Select the page number for the answer to each question in the Gradescope.
  - You may submit typed or handwritten/scanned answers. If you decide to submit handwritten answers then please ensure that it is easily readable.
  - You can easily scan and upload your answers as a PDF using the Gradescope mobile app.

1. **[10pts]** Suppose the numbers of calories in 10 different brands of chocolate milk of 244 mL are: $164, 182, 176, 149, 184, 190, 160, 139, 175, 148$. Assume these numbers are the observed values from a random sample of ten independent normal random variables with mean μ and variance σ^2, both unknown. Find a 95% confidence interval for the mean calories μ.

2. **[10pts]** Let $X_1, \cdots, X_n$ be a random sample from the normal distribution with unknown mean μ and known variance σ^2. How large a random sample must be taken so that 90% confidence interval has length less than 0.02σ?

3. **[10pts]** Consider the setting in question 1, except that we now assume a known variance of $\sigma^2 = 16$. Suppose we wish to test the hypotheses: $H_0 : \mu = 170, H_A : \mu \neq 170$. Determine whether the test rejects H_0 at significance 0.05.

4. **[11pts]** Suppose that nine observations are selected at random from the normal distribution with unknown mean μ and unknown variance σ^2, and for these nine observations it is found that $\overline{X}_n = 20$ and $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 70$. Find p-value of the test with hypotheses: $H_0 : \mu \leq 18, H_A : \mu > 18$.

5. **[10pts]** An experiment is carried out to see if there is any relation between a person's age and whether the person actively uses social media. Suppose that 100 people, 18 years of age or older, are selected at random, and each person is classified according to whether or not they are between 18 and 30 years of age and also according to whether or not they actively use social media. The observed numbers are given in the table below. Test the hypothesis that there is no relationship between a person's age and whether they actively use social media.

	Active social media user	Not active social media user
Between 18 and 30	18	26
Over 30	8	48

Exploratory data analysis on Ted talks data (ted_main.csv)

[**Each part 7 points**] This set of exercises are to familiarize you with the `pandas` library and doing exploratory data analysis with it. The csv file is at https://raw.githubusercontent.com/cpethe/TED_Talks/master/ted_main.csv

6. What is the most common *speaker_occupation* in the dataset?
7. Drop a column from the dataframe that is uninformative (information already contained in other columns), so that the dataframe no longer contains that column.
8. Get the rows corresponding to talks about climate change.
9. Get the rows corresponding to 10 most lengthy talks with at least 10 million views or at least 3000 comments.
10. Which talk is the most viewed as compared to its *related_talks* (the one with the maximum difference between its views and the *view_count* of any of its *related_talks*)?
11. Which pair of features (columns) are the most correlated? Comment on whether the correlation implies causation in this case with a few sentences.
12. Research a way to find out the significance (in terms of p-value) of the correlation of a pair of features. Try it on the pair of columns *duration* and *comments*. What is the correlation coefficient and its p-value? Comment on what this finding implies with a few sentences.