Question 1.a. Estimate a simple linear demand equation by regressing the quantity of gas `quantgas` consumed on the price of a gallon of gas `pricegas`. What is your estimate of the price coefficient from the OLS estimation? Remember to use robust standard errors, and to always include a constant.

```
In [ ]: ...
```

Question 1.b. Use your OLSEs to express the price elasticity of demand evaluated at the average price of gas. Does it make economic sense?

Hint: Express the price elasticity when demand is linear.

Type your answer here, replacing this text.

Question 1.c. Now introduce per capita personal income `persincome` as a regressor in the linear demand model and re-estimate using OLS. How has your estimate of price coefficient changed?

This question is for your code, the next is for your explanation.

```
In [ ]: ...
```

Question 1.d. Explain.

Type your answer here, replacing this text.

Question 1.e. Do you think that the above regression suffers from omitted variable bias? If so, can you determine the sign of the bias?

Type your answer here, replacing this text.

Question 1.f. Give reasons why you should suspect that the gasoline price would be correlated with error term even after you introduced personal income into the regression. Evaluate the monthly sales of autos in the U.S. (carsales) serve as a good instrument for price of gas? Explain.

Type your answer here, replacing this text.

Question 1.g. Estimate the first stage of a two stage least squares estimation by regressing price of gasoline on the sales of cars. Also include personal income. Perform a test that determines whether car sales is a "strong instrument."

This question is for your code, the next is for your explanation.

```
In [ ]: ...
```

Question 1.h. Explain.

Type your answer here, replacing this text.

Question 1.i. Can you suggest another instrument that is likely to be a better instrument than car sales?

Type your answer here, replacing this text.

Question 1.j. Now perform the second stage of the TSLS estimation and report any change in the size of the coefficient on gasoline price as a result of using the instrumental variable.

Hint: ***`results.fittedvalues`*** *will give you an array of the $\hat{y}$ values.*

This question is for your code, the next is for your explanation.

```
In [ ]: gas['pricegas_hat'] = results_1g.fittedvalues
        ...
```

Question 1.k. Explain.

Type your answer here, replacing this text.

Question 1.l. Is the TSLS estimate of the price coefficient statistically significant? Do you have any reason to doubt the reported values of the standard errors from the second stage? Explain.

Type your answer here, replacing this text.

Question 1.m. Suppose you were instead interested in studying how the supply of gas is influenced by its price. Would you feel comfortable regressing the quantity of gas produced on its price? Why?

Type your answer here, replacing this text.

Question 1.n. Also included in the dataset is the BLS monthly price index for consumer purchases of "transportation services" over the same sample period `transindex`. Perform TSLS estimation using this price index as an instrument. Evaluate the results of the first and second stages.

This question is for your code, the next is for your explanation.

```
In [ ]: ...
```

Question 1.o. Explain.

Type your answer here, replacing this text.

Question 1.p. Assume that you are told that at least one of the instruments above is not exogenous (it could be both). Based on your empirical results using these data, decide what you consider the "best" estimate of the price coefficient. It doesn't have to be one of the above instruments. Explain your reasoning.

Type your answer here, replacing this text.

Question 2.a. What percentage of employees volunteered to participate in the experiment?

Hint: Check out the `Series.value_counts()` function.

```
In [ ]: ...
```

Question 2.b.i. Use the variables `commute` as a dependent variable in a bivariate linear regression where `volunteer` is the explanatory variable.

```
In [ ]: ...
```

Question 2.b.ii. Interpret the coefficient on `volunteer` and comment on its statistical significance.

Type your answer here, replacing this text.

Question 2.c.i. Use the variable `tenure` as a dependent variable in a bivariate linear regression where `volunteer` is the explanatory variable.

```
In [ ]: ...
```

Question 2.c.ii. Interpret the coefficient on `volunteer` and comment on its statistical significance.

Type your answer here, replacing this text.

Question 2.d.i. Impressed by your recent econometrics training, Ctrip hires you as a consultant to analyze the results from their experiment. To begin with, you estimate a bivariate linear regression model of the productivity of workers, measured by the log of the average number of calls taken per week (call this variable `ln_calls`), on the variable `WFHShare` (work from home share).

Hint: Add the argument `missing='drop'` when constructing your OLS model to drop the missing entries.

```
In [ ]: ...
```

Question 2.d.ii. Interpret the regression coefficient on `WFHShare` in words. Is the effect statistically significant?

Type your answer here, replacing this text.

Question 2.e. Has the Ctrip company achieved the ideal of a randomized controlled experiement, so that we can view the estimated effects of working from home on productivity in causal terms?

Type your answer here, replacing this text.

Question 2.g.i. Create a dummy variable called `longcommute` which is equal to one if the employee has a commute of greater than or equal to 120 (i.e. 2 hours) and add it to the `ctrip` column.

Hint: First create a boolean column for `longcommute` then cast it into integers using [`Series.astype(int)`].

```
In [ ]: ctrip['longcommute'] = (...).astype(int)
```

Question 2.g.ii. How would you expect that including `longcommute` as a second explanatory variable would alter the coefficient on `WFHShare` – would it increase, decrease, or stay the same? Explain.

Type your answer here, replacing this text.

Question 2.h.i. Management believes that commute (the travel time from home to office and back) is an important determinant of a worker's productivity. They have two hypotheses:

1. Employees who face a longer commute time are generally less productive than workers who have shorter commute times.
2. The effects of `WFHShare` on productivity is larger for those who face a longer commute.

Estimate a regression of `ln_calls`, with `WFHShare`, `longcommute`, and their interaction (call it `WFHShareXlongcommute`) as explanatory variables.

Hint: Once again you will need to add the argument `missing='drop'` when constructing your OLS model to drop the missing entries.

```
In [ ]: ...
```

Question 2.h.ii. Do your results support hypothesis (i), hypothesis (ii), both hypotheses, or neither one? Explain.

Type your answer here, replacing this text.

Question 2.i. If the coefficient on `longcommute` is statistically insignificant, would this lead you to drop `longcommute` from the regression model in part (h)? Explain your answer.

Type your answer here, replacing this text.

Question 2.j. Using the regression in part (h) and without estimating any other regression, write the estimated equation for the simple regression of `ln_calls` on `WFHShare` using only data for those with a commute of fewer than 120 minutes. You must show your solution to obtain full credit.

Type your answer here, replacing this text.

Question 3.a. Treating the ban in cigarette advertising as a quasi-experiment, perform a differences-in-differences analysis of the effect of the ban on the consumption of tobacco. Fill in the table that indicates the conclusion of your analysis.

The top left box with work has been done for you.

```
In [ ]: # Mean of annual grams of Tobacco Sold per Adult (15+) across the pre-treatment periods in Cana
        pre_period = cigads[cigads['YEAR'] <= 1970]
        np.mean(pre_period[pre_period['COUNTRY'] == "CAN"]['CIGSPC'])
```

Question 3.b.i. Now create a dummy variable `post` indicating the time period whether the ban was in effect or not, plus a dummy variable `treat` for the treatment group (i.e. the U.S.) and the control group (i.e. Canada). Regress tobacco consumption on these two dummies and on the interaction between the two (you can call this `treatpost`).

Hint: Once again you will need to first create boolean columns then cast it into integers using *`Series.astype(int)`*.

```
In [ ]: cigads['post'] = (...).astype(int)
        cigads['treat'] = (...).astype(int)
        cigads['treatpost'] = ...
        model_3b = ...
        results_3b = ...
        results_3b.summary()
```

Question 3.b.ii. How do your results compare to your diffs-in-diffs estimator?

Type your answer here, replacing this text.

Question 3.c.i. Finally, recognizing that price does also affect consumption, you introduce the price variable into the regression in (b).

```
In [ ]: model_3c = ...
        results_3c = ...
        results_3c.summary()
```

Question 3.c.ii. Report your results and compare to those from (b).

Type your answer here, replacing this text.

Question 3.d. Why would you expect that the price of a pack of cigarettes might be correlated with the error term? Note that some economists have argued that the advertising ban reduced competition among cigarette makers by eliminating one dimension on which they compete for customers, which in turn led to higher prices.

Type your answer here, replacing this text.

Question 4.a. Construct the average values of `entercollege`, `hsgpa`, `privatehs`, `hidad`, `himom` for each integer value of `psu` (e.g., get the averages for scores from 300 to 300.99, and assign them to the "300" bucket; then get the averages for scores from 301 to 301.99 and assign them to the "301" bucket, etc.). This is sometimes called "collapsing" the data to integer cells. This is a bit tricky, so we provide the commands for you below.

```
In [ ]: rd['psu_integer'] = np.floor(rd['psu'])
        rd_temp = rd.groupby('psu_integer').agg(['mean']).reset_index()

        rd_collapsed = pd.DataFrame()
        rd_collapsed['psu_integer'] = rd_temp['psu_integer']
        rd_collapsed['hsgpa'] = rd_temp['hsgpa']['mean']
        rd_collapsed['psu'] = rd_temp['psu']['mean']
        rd_collapsed['entercollege'] = rd_temp['entercollege']['mean']
        rd_collapsed['privatehs'] = rd_temp['privatehs']['mean']
        rd_collapsed['hidad'] = rd_temp['hidad']['mean']
        rd_collapsed['himom'] = rd_temp['himom']['mean']
        rd_collapsed['over475'] = rd_temp['over475']['mean']
        rd_collapsed.head()
```

Question 4.b. Generate plots of the average values of `entercollege`, `hsgpa`, `privatehs`, `hidad`, `himom` (from 4.a) as a function of `psu` (be sure to label your axes and give each plot a title). You should see a jump in `entercollege` at 475 points, but relatively smooth values of the other variables. The following cell is for your code.

```
In [ ]: plt.scatter(...)
```

Question 4.c. Next you will fit *local linear* regressions using different bandwidths. To do this you will regress one of the dependent variables Y_i on the following independent variables: `constant`, `psu`,`over475` and $\tilde{psu} = psu - 475$, i.e., you will fit the model

$$Y_i = \beta_0 + \beta_1 over475_i + \beta_2 \tilde{psu}_i + \delta_3 (\tilde{psu}_i \cdot over475_i) + u_i$$

Interpret the coefficients of this regression model.

Type your answer here, replacing this text.

Question 4.d. Using the "collapsed" data from part 4.a, which has one observation per integer value of `psu_integer`, and a bandwidth of 10 on each side of the 475 cutoff, fit the model for each of the dependent variables $Y_i = entercollege$, $Y_i = hsgpa,$, $Y_i = hidad$, $Y_i = himom$ (i.e., you are fitting four separate models here). The following cell is for your code.

Hint: This means that you fit the regression models to the collapsed data for the subset of data with $465 \leq psu\_integer \leq 485$*. This data set will have 21 observations – 10 observations for scores less than 475 and 11 observations for scores of 475 or higher.*

```
In [ ]:
```

Question 4.e. Repeat part 4.d using a bandwidth of 20 points. Do you find that the estimated jumps are similar for all four dependent variables as with a bandwidth of 10?

The first cell is for your code, the second cell is for your question answer.

```
In [ ]:
```

Type your answer here, replacing this text.

Question 4.f. For every bandwidth from 5 to 50, develop a plot to show the estimate of β_1 when the dependent variable Y_i is *entercollege*. The following cell is for your code.

```
In [ ]:
```