**Assignment 1)**

**Part 1: Feature Engineering**

The file "[nyc-rolling-sales\_clean.csvPreview the document](https://canvas.uchicago.edu/courses/25628/files/3012289/download)" contains New York City property sales data [taken from Kaggle (Links to an external site.)](https://www.kaggle.com/new-york-city/nyc-property-sales). The data has already been cleaned and a few columns have been dropped.

Write your own code in Python that performs the following operations on this dataset:

* Create a new column "year of sale" by extracting year information from "SALE DATE" column.
* Drop the "SALE DATE" column.
* Get dummy encoded columns for 'TAX CLASS AT PRESENT' predictor.

**Part 2: Hypothesis Testing**

The file "[gestational\_study.csvPreview the document](https://canvas.uchicago.edu/courses/25628/files/3012291/download?wrap=1)" contains data from a small study involving 17 infants on the relationship between gestational age at birth (measured in weeks) and birth weight (measured in grams).

Write your own code in Python that performs the following operations on this dataset:

* Calculate the correlation between Gestational Age and Birth Weight (use the corr() function in Pandas to calculate the correlation).
* Generate a scatter plot to visualize the relationships between these two variables. Does the scatter plot confirm the correlation value you calculated? Does there seem to be a strong correlation between the two predictors?
* Perform a hypothesis test and state the confidence level at which you can confirm a correlation between the two variables (use the scipy.stats.pearsonr() function--consult [the SciPy documentation for details (Links to an external site.)](https://docs.scipy.org/doc/scipy/reference/index.html) about its inputs and outputs).