Question 1 (16 points)

A logistics company wants to analyze the fuel consumption of their fleet of vehicles. For every trip, they record the vehicles distance travelled, average speed, trip duration, and fuel consumption as well as the average temperature during the trip. They want to know how they can use this data to predict their fuel consumption in the future.

a. Load and format the data (1 point)
b. Search for collinearity. Based on your finding, select a subset of columns to use in part C. (5 points)
c. Perform a linear regression after removing the collinearity (2 points)
d. Fit an appropriate distribution to each of your independent variables. The distribution should fit the data, but you do ***not*** need to prove the best fit (*i.e.* using loglikelihood) Using these distributions and the linear regression found in part c, create 1000 samples of fuel use for individual trips. Plot a histogram of the sampled values and compare to the original data. (8 points)

Question 2 (16 points)

Babies are weighed frequently through their first three years of life to ensure they are developing at a suitable rate. Deviations from the usual growth curve could indicate a serious health concern that requires early intervention.

The sample data for this question includes measurements of a child's weight at one-month intervals. Each column represents a time of measurement, each row represents an individual. Using this data, perform the following:

a. Load the data and convert it to a data frame with columns for index (individual identifier), age, and weight (2 points)
b. Fit a linear mixed effects model using age and an independent variable, weight as a dependent variable, and individual as the random variable. Plot the predicted mean values at ages 0 to 36 months using this model to the mean observed values. Describe the fit. (10 points)
c. Transform the independent variable (age) using a function that you think will improve the model's predictive ability. Fit a second linear mixed effects model using this transformed variable. Using metrics and figures, compare the fit of both models. (4 points)

Question 3 (12 points)

An industrial company is assessing the possibility of installing solar panels to offset some of their energy costs. For the past week, they have recorded the generation from a small test panel located at their site as well as their hourly energy use.

a. Load the data and convert it so that you can fit a B-Spline with the hour-in-day (*e.g.* 1pm) as the x axis. Remember that scipy.interpolate expects the x values to be sorted. (3 points)
b. Fit a B-Spline of degree 3 to the generation and load data. You will need to specify a smoothing parameter s to avoid numerical errors. Plot the original data with the B-Spline (3 points)
c. Adjust the adjusted hour-in-day using a cosine transform such that it has a value of 1 at midday and 0 at midnight. Use this transformed data to create truncated power function of degree 2

with a knot at 0.5. Fit the generation data to this function using a linear regression. Plot the original data against this curve (4 points)

d. Compare the fits of the generation data you created in part (b) and part (c). Which method do you prefer, justify your answer. (2 points)