STAT 5428: Homework 1

Consider the "Pima Indians Diabetes Database" (https://www.kaggle.com/uciml/pima-indians-diabetes-database). This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage. The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables include: Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function and Age. The data includes 768 observations.

Please do the following:

1. Calculate the mean, median, standard deviation, IQR, skewness, and kurtosis for all the variables.
2. For each variable, construct a histogram with 5 bins and comment on the shape of the distribution. Identify the variables that have similar histograms in terms of the shape. Repeat this process with 10 bins and 20 bins. What do you observe?
3. Construct box-and-whisker plots for all variables.
4. Construct side-by-side boxplots for all the variables for the Outcome=1 and Outcome=0 groups. Identify the variables that have boxplots with different shapes between the two groups? What does it mean?

Please note that your results and comments should be provided in one single pdf file. Please try to use IJulia notebooks to submit the homeworks. Incase if you want try different, you may use markdown to generate a report.