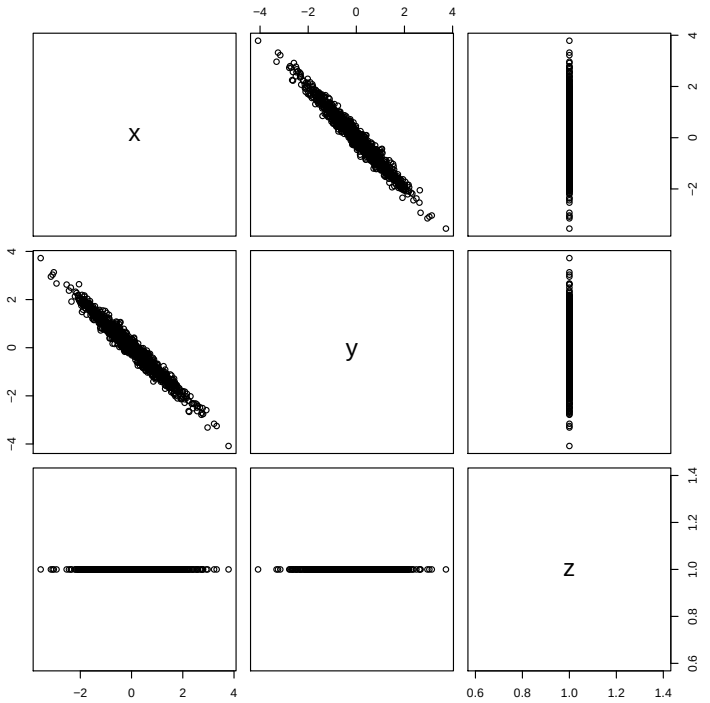
B565: Homework 3

1. The figure below describes the scatter of 1000 examples of 3 different variables, x, y, z.



   From the figure approximate the diagonalization of the covariance matrix $S = UDU^t$ by writing down the U and D matrices.

2. Consider the "Arrests.csv" data on Canvas.

   (a) Compute estimates of the probability of being released for "Black" and "White" arrested individuals, and plot this information in a mosaic plot.

   (b) Variables x and y are *conditionally independent* given z if $p(x, y|z) = p(x|z)p(y|z)$. This definition is analogous to the usual definition of independent variables, only now the independence is under the conditional distribution on z. We can also describe conditional independence as $p(x|y, z) = p(x|z)$ — the conditional distribution $p(x|y, z)$ does not *depend* on y. This is analogous to the alternative characterization of independence we developed in class: for independent variables the conditional distribution doesn't depend on the variable we condition on. If the variables *Released* and *Colour* were conditionally indepdent given *Employment* that would suggest that the higher arrest rate among "Blacks" may be illusory, with the real "cause" being *Employment*.

   Plot the mosaic plots of *Colour* and *Released* separately from the employed and not-employed individuals. Does it appear that *Colour* and *Released* are conditional independent given *Employment*? Explain your reasoning carefully.

   (c) The *Checks* variable counts the number of various databases that identify the individual — something like a measure of previous contact with the polic department. Does it appear that, given the number of *Checks* that *Released* and *Colour* are independent?

   (d) Do the data appear consistent with racial bias? Explain your answer carefully discussing what can be known and what cannot be inferred from these data.

3. Create a random sample $x_1, \ldots, x_n$ for $n = 1000$ from the Exponential(1) distribution using the R command "rexp" (analogous to runif and rnorm).

(a) The cumulative distribution function (cdf) of a random variable, X, is defined to be $F(x) = P(X \leq x)$. So, using the percentile notation discussed in class $F(x_{50}) = .5$, $F(x_{25}) = .25$, etc. Create a plot of the empirical cdf for this sample as follows. For each x_i in your sample, plot a point (x_i, y_i) where y_i is the empirical estimate of the probability of being less than or equal to x_i. That is, y_i is the proportion of your samples that were $\leq x$. Show the graph with a smooth curve, rather than an a sequence of points.

(b) Create a new data set by transforming each point by F where F is the cdf you computed in the previous part. That is, your new data are $F(x_1), \ldots, F(x_n)$. What is the distribution of these new samples?

(c) Suppose you do the same experiment using random numbers from a different distribution. Argue that the result of the *transformed* points will still have the same distribution as in the previous part.

4. The file "time_series.csv" on the Canvas site contains 200 time series generated from one of four different models. Each time series (each row of the .csv file) has length 200. The models are unknown to you and are chosen randomly for each time series.

   (a) Visualize these different time series in a way that allow you distinguish the 4 models. Submit a plot that illustrates the fundamental ways in which the 4 categories differ. Characterize this difference in words.

   (b) Derive two features that effectively separate the time series into four categories when shown in a scatterplot. There may by many ways to do this. Submit your scatterplot as well as the code that generates the features.

5. Suppose a pair of critics rate a collection of $N = 1000$ movies giving a numerical score for each movie. A reviewer cannot give two movies the same score. One critic is generally more positive than the other, so these scores are reported as *percentiles*. That is, each movie gets a score in $[0, 1]$ from each reviewer which is the overall fraction she judges no better than the current movie. Denote the reviewers' percentile scores as x_i, y_i, $i = 1, \ldots, N$.

   (a) Sketch a reasonable scatterplot for the the pair (x_i, y_i)

   (b) Would it be reasonable to model the two reviewers scores as independent random variables? Why or why not?

   (c) Is it possible that one reviewer always gives a higher percentile score than the other reviewer? Either give an example showing this is possible or argue that it is impossible.

   (d) Is it possible that difference in percentile scores, $x_i - y_i$, is, on average, greater than 0? Either give an example showing this is possible or argue that it is impossible.

   (e) Is is possible that $x_i > y_i$ for all but 1 movie? Either give an example showing this is possible or argue that it is impossible.

6. Suppose you are given an $n \times n$ matrix of positive numbers, D. Is it always possible to find a collection of n points (x_i, y_i), $i = 1, \ldots, n$ so that the Euclidean distance between (x_i, y_i) and (x_j, y_j) is D_{ij}? Why or why not?

7. In your book, problem 3.10.