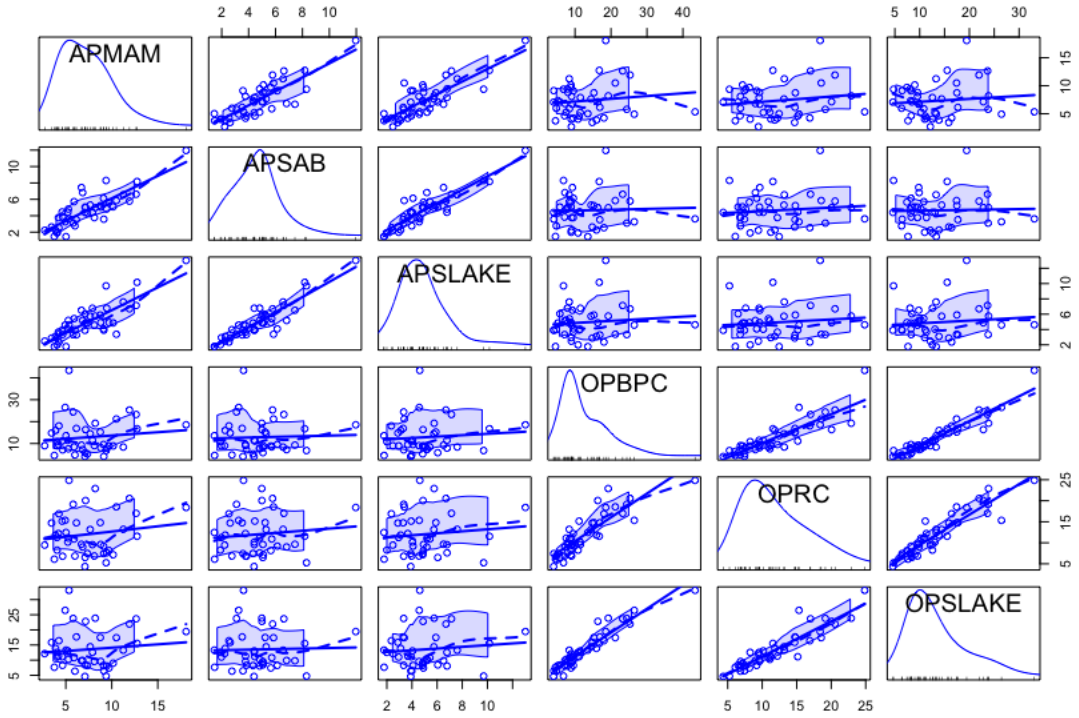
Exercise 1.5: Can southern california's water supply in future years be predicted from past data? One factor affecting water availability is stream runoff. If runoff could be predicted engineers, planners, and policy makers could do their jobs more efficiently. The data file contains 43 years work of precipitation measurements taken at six sites in Sierra Nevada mountains and stream runoff volume at a site near Bishop, California. Draw a scatter plot for these data and summarize the information from these plots.

Solution:
Loading in our data and creating a scatter plot matrix we get the following visualization,

Figure 1: Correlation Scatter Matrix for SRV



We can see that there is a strong positive correlation between the OPSLAKE, OPRC, and OPBC sites. There is also a strong positive correlation between the APSLAKE, APSAB, and APMAM sites. The rest of the scatter plots exhibit high variance and little to no correlation. This is likely due to the fact that stream runoff was measured in 2 places, the Sierra Nevada Mountains, and Bishop, California. All plots seem to be positively skewed from the presence of an outlier. Looking into the data it seems like 1962 was a strong year for runoff in Bishop, California and 1982 was a strong year in the Sierra Nevada Mountains.

Code:

```
> scatterplotMatrix(df[c(1:43),c(2:7)])
```

Exercise 2: For the data set in problem 1.5, do the following

1. Give at least two possible ways in which the data set could be inadequate for building a useful prediction model for stream runoff near Bishop, California.

   Solution:
   While data from all sites looks to have a close constant variance its clear that in general the variance is relatively very high and we would be hard pressed to create a predictive linear model that is accurate.

   In general all the data is positively skewed with the OPBC and OPSLAKE sites in particular having very high outliers. Not addressing these extreme values can cause a linear model to be less accurate. **Code:**

```
> plot(df[c(1:43), c(1)],df[c(1:43), c(7)] )
> plot(df[c(1:43), c(1)],df[c(1:43), c(6)] )
> plot(df[c(1:43), c(1)],df[c(1:43), c(5)] )
> boxplot(df[c(1:43), c(7)] )
> boxplot(df[c(1:43), c(6)] )
> boxplot(df[c(1:43), c(5)] )
```

2. Does the data set result form an observational, or experimental, study?

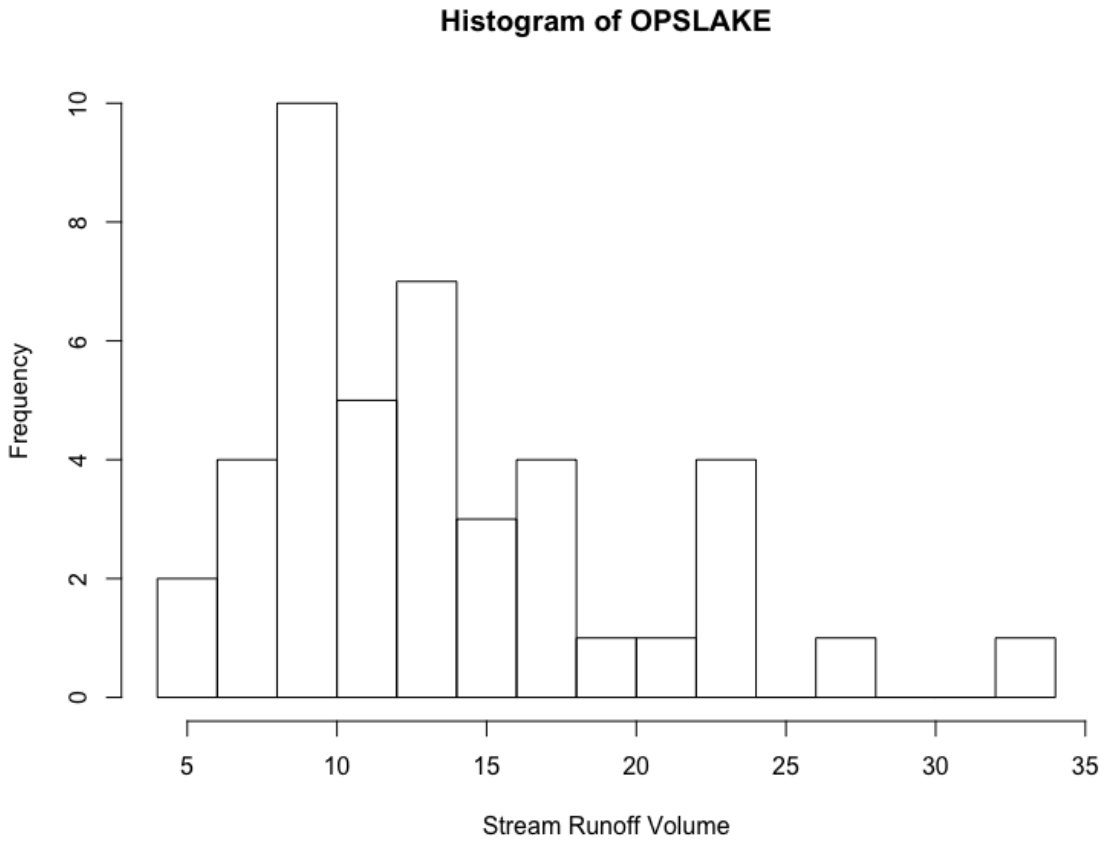
   Solution:
   Since the researchers weren't purposefully manipulating some variable in an effort to measure a response, this data is observational data. The goal is to collect data without interfering.

3. Give at least three important questions that a person might have about the data set before performing inferential analysis on it.

   Solution:
   Primarily we want to know the Sampling Method used to recover the data. This will influence how we estimate the parameters we want to know. We also want to now how dirty or clean the data is, if there were issues or changes in how the data was recovered, those things need to be taken into account during our analysis. We also want to have some idea of how the data is distributed(Can we assume normality?)

4. Create some kind of plot or visualization to answer one of the questions you listed in the last part.

Solution:
The following is a histogram of the runoff volume in the OPSLAKE site. This gives us an idea of the how the data is distributed. It seems like it might be normal with a slight positive skew because of the outlier value. We might benefit from bootstrapping this data.



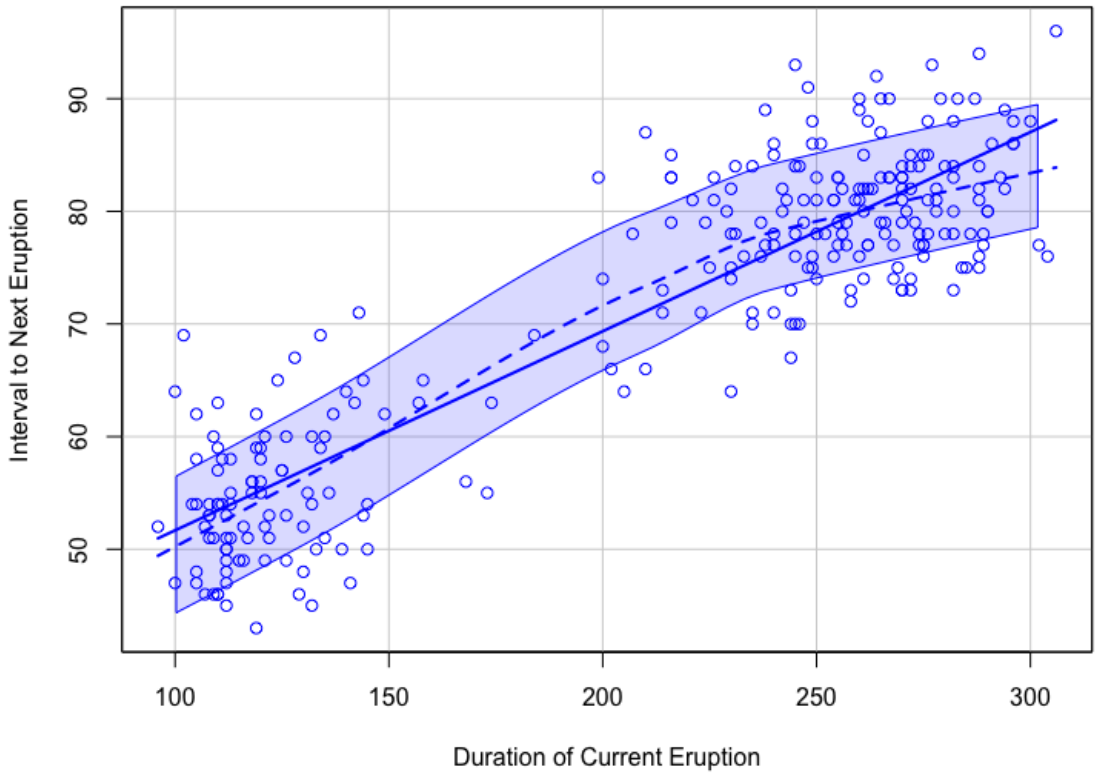
Code:

```
hist(df[c(1:43), c(7)], breaks = 15,
     main = 'Histogram of OPSLAKE',
     xlab = 'Stream Runoff Volume')
```

Exercise 1.4: The data file gives information about eruptions of Old Faithful Geyser during October 1980. variables are the *Duration* in seconds of the current eruptions, and the *Interval*, the time in minutes to the next eruption, The data were collected by volunteers and were provided by the late Roderick Hutchinson. Apart form missing data for the period from midnight to 6 a.m. This is a complete record of eruptions for that month. Draw the relevant summary graph for predicting interval from duration and summarize your results.

Solution:
Importing the data and plotting with r, we get the following scatterplot,

Figure 2: Duration vs. Interval ScatterPlot



Looking at the scatter plot, beyond the positive correlation between duration and interval the data appears to be bimodal. There seems to be a point where if the duration is around 200 seconds or more the interval to the next eruption is greater than 70 while if the duration is less than 200 the interval for the next eruption is less than 70.

Exercise 4: For the data set in problem 1.4 do the following,

1. Identify the predictor and response.

   Solution:
   In this data set we use the duration is the predictor variable and the interval is the response variable.

2. Does the data enable you to make claims about causation, or merely association?

   Solution:
   This is another example of data that is observational, since we had no role in changing the duration variable, or even developing a control duration variable. As a result we can only make claims about association.

3. Does the straight-line mean function seems to be plausible for this data set? Why or why not?

   Solution:
   For the most part the variability seems to be constant, I would be worried around the 200 seconds interval but I think more data would need to be collected to say for sure.

4. Give the mean function and variance function (in terms of β_0, β_1 and σ^2) that would comprise a linear model for this data set.

   Solution:
   Using the lm function in r we get that,

   $$E(Interval|Duration = x) = B_0 + B_1 x = 33.988 + .177x$$

   $$Var(Interval|Duration = x) = \sigma^2 = (1.182)^2$$

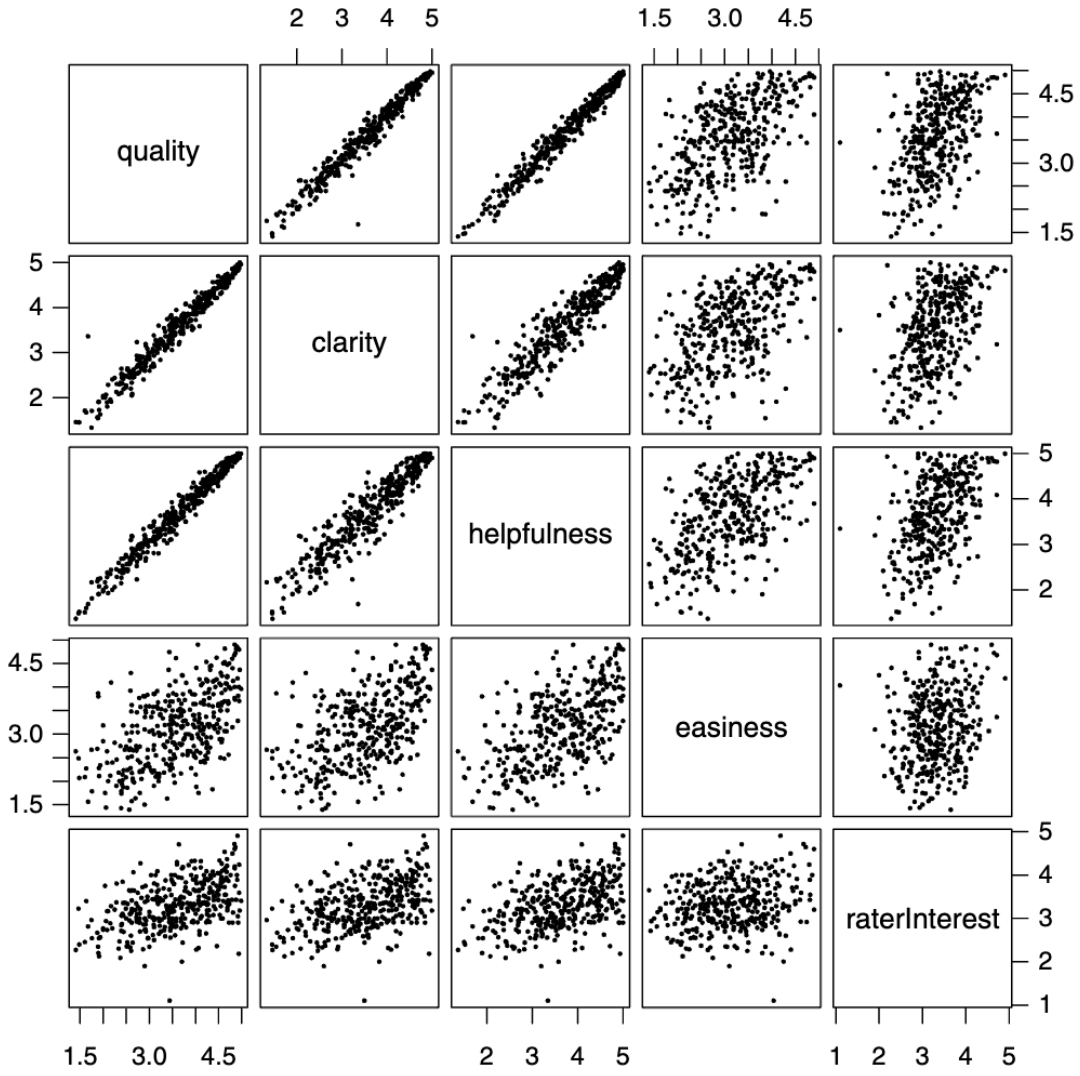
   Code:

```
> lm1 = lm(Interval~Duration, data = df)
> summary(lm1)

Call:
lm(formula = Interval ~ Duration, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
−12.3337   −4.5250    0.0612    3.7683   16.9722

Coefficients:
             Estimate  Std. Error  t value
(Intercept) 33.987808    1.181217    28.77
Duration     0.176863    0.005352    33.05
```

Exercise 1.6: Provide a brief description between the five ratings described in the data from ratemyprofessor.com.

Figure 3: Average professor ratings from ratemyprofessor.com



Solution:
From the plot we can see that there is a strong positive correlation between the quality, clarity and helpfulness of the professor ratings. Furthermore it seems like easiness and raterinterest are relatively uncorrelated with the rest of the data which might suggest that in aggregate the ratings are unbiased.