1. The following data step reads in data related to a study of beta carotene in individual's diets. The meaning of id, age and sex are apparent, smkstat is smoking status (a categorical variable with numerical values) and dietcar1 to dietcar5 represent measures of dietary intake for the individual over 5 years.

```sas
data a;
input id age sex smkstat dietcar1 dietcar2 dietcar3 dietcar4 dietcar5;
run;
```

(a) Specify what SAS code you would add to the data step if you wanted to create two new variables, avgdiet = average of dietcar1 to dietcar5 and ranged = maximum of the five diet values - minimum of the five diet values.

Write out the SAS code to accomplish each of the following:

(b) List the data but only showing age, sex and the 5 diet measures.

(c) Create a permanent SAS data set on the a: directory.

(d) Create a temporary/internal SAS data set called b, which has only those cases with sex=1.

(e) Gets a complete set of descriptive statistics for the variable dietcar1 including i) stem and leaf plot and box-plot ii) normal probability plot and tests for normality iii)90% confidence intervals for the population mean and standard deviation under normality iv) confidence intervals for the population percentiles/quantiles under normality v) confidence intervals for the population percentiles/quantiles without assuming normality. State which part of your code is responsible for the desired results for i) - v).

(f) Obtains the number of individuals in each category of smoking status and provides a bar chart of these frequencies.

(g) Obtains both the Pearson and Spearman correlation between dietcar1 and dietcar5.

(h) **Within each level of smoking status gets a 95% confidence interval for the difference between dietcar5 and dietcar1 and tests the null hypothesis that the means of these two variables are equal.**

(i) Compares avgdiet (created in part a)) between the two sexes via a t-test that the two means are equal and finds a 99% confidence interval for the difference between means.

(j) Creates a table showing the frequencies for each combination of sex and smkstat.

(k) Plots dietcar1 versus age for each sex.

(l) Estimates the proportion of people with dietcar5 greater than dietcar1 and finds an approximate confidence interval for the proportion. To do this you will first have to add some code to the data step.

(m) Write a data step which creates a new SAS data set which has in it variables id, sex, year, diet (and only those variables) where year goes from 1 to 5 and diet has the corresponding diet value for that year. So each initial record now produces 5 records.

2. The data below are from an 1854 survey conducted by the Massachusetts Commission on Lunacy under the leadership of Edward Jarvis. Dr. Jarvis was President of the American Statistical Association from 1852 to 1882. Write an input statement that will read in the data being sure that you get the full county name and account for any missing values.

```
12345678901234567890123456
COUNTY     NBR   DIST
```
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BERKSHIRE</td>
<td>119</td>
</tr>
<tr>
<td>FRANKLIN</td>
<td>84</td>
</tr>
<tr>
<td>HAMPshire</td>
<td>94</td>
</tr>
<tr>
<td>HAMPDEN</td>
<td>105</td>
</tr>
<tr>
<td>WORCESTER</td>
<td>351</td>
</tr>
<tr>
<td>MIDDLESEX</td>
<td>357</td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
</tbody>
</table>