1. A SAS data set was created using the following input statement:

```
input state $ city $ pop1994 income housing electric;
```

Answer parts (a) to (f) below.

(a) (2) Give the name of a SAS procedure (that is not a graphics procedure) that allows you to construct a normal probability plot of income and add an inset to the plot.

```
univariate
```

(b) (2) Give the name of a SAS procedure that you may use to obtain a scatterplot matrix of variables pop1994, income, housing, electric using ODS statistical graphics.

```
gscatter
```

(c) (2) Give the name of a SAS procedure that may be used to examine the distribution of variables such as housing by plotting histograms overlaid with density plots using ODS statistical graphics.

```
gplot
```

(d) (2) Give the name of a SAS/GRAPH procedure that you may use to fit a linear regression line (and also add prediction intervals) to interpolate the data points in a scatter plot of electric vs. pop1994 in high resolution graphics.

```
gplot
```

(e) (2) Give the name of a SAS/GRAPH procedure that you may use to plot horizontal bar graphs that shows the averages of electricity usage and group them by a classification variable (such as region) in high resolution graphics.

```
gchart
```

2. (4) Use the following plot to say in what way the distribution of the data is different from that of a normal distribution:

```
The bowl shape of this plot indicates that the distribution is right-skewed compared to the normal distribution.
```
3. (a) What is the plot used for comparing a sample distribution to a theoretical distribution using percentiles of each distribution? **QQ plot**

(b) The strength of the linear relationship between two variables in a bivariate sample is indicated by the sample correlation coefficient. What plot should you use to verify this relationship? **Scatter plot of the two variables**

(c) Name three graphical methods (plots) available for studying the distribution of a (univariate) random sample. **Histogram, box plots, quantile plot**

(d) Name a plot that can be used to examine relationships among the variables in multiple regression. **Scatter plot matrix**

4. The graph below of side-by-side boxplots summarize life expectancies of 60 countries according to the category variable level of technology.

(a) Compare the shapes of the distributions for the three categories of level of technology. Estimate largest life expectancy for all countries? The smallest? **Shapes change from right-skewed, symmetric to being left-skewed. Largest ~ 76.5, Smallest ~ 43.5**

(b) Compare the location of the distributions for the three categories of level of technology. Which category appears to have more countries with higher life expectancy? **Medians increase from 51, 59 to 73.5. High tech countries have higher life expectancy, too.**

(c) Compare the spread (or dispersion) of the distributions for the three categories of level of technology. Which category of countries (with respect to level of technology) have more countries with a large range of life expectancy values? **Small dispersion to high and then to moderate dispersion. The moderate technology level countries have a larger spread of life expectancy.**

(d) Name 3 features in the boxplots below you could use to verify that model assumptions made when analyzing this data using ANOVA methods are plausible. **Distribution: is it normal? Spread: are they constant across levels of technology? Are there any outliers?**

Box Plots of Life Expectancy by Level of Technology
5. Below is a Q-Q plot of measurements of tire wear on two different brands of tires:

(a) Using the above graph, compare the shapes of the distributions of tire wear measurements of the two brands of tires. The two tire-wear distributions have the same shape.

(b) Using the above graph, compare the relative magnitudes of tire wear of the two brands of tires.
\[
\text{Median}_1 = 16.2 \quad \text{Median}_2 = 12 \quad \text{slope} = \frac{12}{10} = 1.2 \quad \Rightarrow \quad \frac{\text{Var}_1}{\text{Var}_2} = 1.2^2 = 1.44 \quad \text{Var}_1 = 1.5 \times \text{Var}_2
\]

6. Below is a Quantile plot of the cities data set used in a textbook example:

Estimate, as accurately as possible, the following quantities using the graph:

(a) 40th percentile \( \sim 16,500 \)

(b) Inter-quartile Range \( \sim 23,000 - 15,000 = 8,000 \)

(c) Observe that about 60% of the cities in this data set has values of median housing costs in a very narrow range. Estimate the narrowest such range. The part between .1 and .7 is the flattest. The range of median housing cost is \( 12,000 - 21,000 \)

(d) What percentage of cities have a median housing cost above $27,000 approximately? \( 10\% \)
7. Consider following data:

<table>
<thead>
<tr>
<th>Case#</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>26.5</td>
<td>18.3</td>
<td>21.8</td>
<td>19.6</td>
<td>19.4</td>
<td>30.7</td>
<td>20.2</td>
<td>16.8</td>
<td>16.9</td>
<td>14.0</td>
<td>15.4</td>
<td>15.2</td>
</tr>
</tbody>
</table>

The procedure `reg` in SAS was used to perform analysis of this set of data using the model

\[ y = \beta_0 + \beta_1 x + \epsilon \]

where \( \epsilon_i \) are assumed to be independently distributed as \( N(0, \sigma^2) \) variables. Answer the following questions based on the results appearing in the output attached to the end of the question.

i) (2) What is the value of \( R^2 \)? What does this say generally about the fit of this model?

\[ R^2 = 44.42 \text{ or } 44.42\% \]

The fit of the model is weak to moderate.

ii) (2) Give the residual sum of squares and its degrees of freedom. What is the estimate \( \hat{\sigma}^2 \) of \( \sigma^2 \)?

\[ SSE = 348.47615 \text{ with 10 d.f.} \]

\[ \hat{\sigma}^2 = \frac{SSE}{\text{d.f.}} = 34.84762 \]

iii) (2) What is the F statistic for testing \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \)? Make a decision based on the p-value.

\[ F = 7.01 \quad p\text{-value} = 0.0179 \]

\[ p\text{-value} < \alpha = 0.05 \]

Reject \( H_0 \) at \( \alpha = 0.05 \)

iv) (4) Using the estimate of \( \beta_1 \) and its standard error, compute the t-statistic for testing \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \).

\[ t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{-1.03259}{0.36524} = -2.8212 \]

v) (4) Using the estimate of \( \beta_1 \) and its standard error, compute a 95% confidence interval for \( \beta_1 \).

\[ \hat{\beta}_1 \pm t_{0.05, 10} \times \text{SE}(\hat{\beta}_1) \]

\[ -1.03259 \pm 2.18 \times 0.36524 \Rightarrow (-1.7963, -0.2687) \]

vi) (4) Use the value of \( h_{33} \) to compute the standard error of the residual for observation 3.

\[ \text{SE}(e_3) = \sqrt{\frac{e_3^2}{1 - h_{33}}} = \sqrt{34.84762 \times 0.4024} \]

\[ = 5.5928 \]

vii) (4) Use the value of \( h_{33} \) to compute the standard error of \( \hat{y}_3 \).

\[ \text{SE}(\hat{y}_3) = \sqrt{\frac{1}{h_{33}}} \text{SE}(\hat{y}_3) = 34.84762 \times \sqrt{0.4024} \]

\[ = 1.8890 \]
viii) (4) Use the the residual for observation 3 and its standard error to calculate the corresponding studentized residual.

\[ t_3 = \frac{e_3}{\text{SE}(e_3)} = \frac{-2.9239}{5.5928} = -0.5228 \]

ix) (4) Use an appropriate plot to determine possible y-outliers. Name this plot, give the case number(s) of the suspected y-outliers and explain why you think each is a y-outlier.

Plot B or C
Case #6 & #12
Both outside cut-off value of \( \pm 2 \) \( \Rightarrow \) possible y-outliers

x) (4) Find any cases, if any, that may be x-outliers using a suitable cut-off value. Name the plot that you may use to determine if x-outliers are present. Does this plot indicate any x-outliers?

Cut-off for hat’s = \( \frac{3.2}{n} = \frac{4}{12} = 0.333 \)
Case #6 \( \hat{h}_{66} = 5.5785 \) exceeds 0.333 \( \Rightarrow \) x-outlier
Plot E shows Case #6 outside the cut-off

xi) (4) Find any cases from the output statistics (give case number), if any, that may be influential explaining why you selected these. Use a plot to determine an influential case, name this plot and say why you selected this case.

Cut-off for Cook’s D = \( \frac{4}{n} = 0.333 \)
Cook’s D for Case #6 = 2.5 exceeds above & is clearly an influential case.
Plot E shows Case #6 far above this cut-off.

xii) (4) If you find any case to be influential, explain why or why not this case should be examined carefully. Use other related case statistics and plots in your explanation.

Case #6 is highly influential, is an x-outlier but more importantly a possible y-outlier. Thus this case may not fit the model and thus must be checked.

xiii) (4) Examine the residuals vs. predicted values plot. State what you think this plot indicates about the fit of the data to a straight line model. Is the straight line model adequate?

Plot A does show that the spread seeming to increase as \( \hat{y} \) increases; but this may be attributed to the y-outliers that are present.

xiv) (4) Examine the normal probability plot of the residuals. Name this plot. What assumption about the model is checked using this plot? Explain whether the plot provides evidence to support this assumption.

Plot I, checks the validity of the normality assumption of errors, \( (e_i's) \) show no pattern or deviation from a straight line, thus supporting this requirement.