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

```
input St $ Pop Tax Numlic Income Roads Fuel;
```

Answer parts(a) to (e) below. (The answers to some of these parts may be statistical graphics (SG) procedures.)

(a) (2) Give the name of a SAS procedure that you used in class to test the hypothesis \( H_0 : \mu = 4 \) vs. \( H_a : \mu \neq 4 \) concerning the mean of variable Income (per capita income, measured in thousands of dollars).  **_univariate_**

(b) (2) Give the name of a SAS procedure that produces one-way or two-way tables of frequency counts and \( \chi^2 \) tests of independence for category variables TaxGrp and IncomGrp.  **_freq_**

(c) (2) Give the name of a SAS procedure that enables you to create a variety tables of statistics computed for variables such as Fuel, classified by variables such as LicGrp and IncomGrp, that may be highly customized.  **_tabulate_**

(d) (2) Give the name of a SAS procedure to produce horizontal barcharts of fuel use for three income groups in separate panels showing data for States in two fuel tax groups.  **_sgraph_**

(e) (2) What is the SAS procedure that you can use the Shapiro-Wilk to test for normality of the distribution of the Roads variable above?  **_univariate_**

2. (a) (3) Examine the proc step

```
proc format;
  value rf 1='Fast'
            2='Moderate'
            3='Slow' ;
run;
```

Suppose that the variable RunGrp (a category variable defined using the variable RunTime used in class) has values of 1, 2, or 3. Write a SAS statement would make SAS print the formatted values of this variable in a proc step used to analysis of the corresponding data set.  **_format RunGrp rf.;_**

(b) (3) To examine whether a Poisson distribution provides a reasonable model for the number of noxious weeds found in in samples of meadow grass, the following statement is included in a proc freq step:

```
tables weeds/nocum testp=(4.9 14.7 22.3 22.4 16.9 10.2 5.1 2.2 1.3);
```

Explain as much as possible the details of the statistical test this statement will produce.  **_Tests whether the observed weed counts fit the Poisson distribution using the X goodness-of-fit test._**

(c) (3) The following statement is included in a proc sgplot step using the fitness data set as input, where OxyGrp and RunGrp are group variables each with 3 categories:

```
hbar OxyGrp/response=Aero stat=mean group=RunGrp;
```

Explain as much as possible the SAS output this statement will produce.  **_Horizontal barchart of means of Aerobic points plotted at each OxyGrp, with each subdivided into groups of Runtime variable._**
(d) (3) The following statement is included in a proc univariate step in a class SAS example using the biology data set as input where Height is a quantitative variable.

```
histogram Height/midpoints=60 to 78 by 3 normal;
```

Explain as much as possible the SAS output this statement will produce.

**Histogram of Height variable with 7 bins each centered at 60, 63 ... etc. overlayed with a normal density fit to the data**

3. Tires of two different brands were driven the same distance in 16 trials and the tire wear measured. Below is a Q-Q plot of measurements of tire wear of the two different brands of tires:

**Q-Q plot of Tire Wear Data**

(a) (2) Using the above graph, compare the shapes of the distributions of tire wear measurements of the two brands of tires.

The tire wear distributions of the two brands have the same shape as the points lie on a straight line, approximately.

(b) (2) Using the above graph, compare the median tire wear of the two brands of tires.

\[
M(\text{Median for both samples}) = \frac{Y_0 + Y_0}{2}
\]

\[
M(\text{Brand #1}) = 16 \quad M(\text{Brand #2}) = 12.5
\]

(c) (2) Using the above graph, compare the variability (as measured by \( \sigma^2 \)) of tire wear of the two brands of tires.

Calculating the approx. slope = \[
\frac{23 \cdot 2 - 12.5}{18 - 16} = 1.34
\]

\[
\frac{\sigma_1^2}{\sigma_2^2} = (1.34)^2 \approx 1.8
\]

Variability of Brand #2 Tirewear is 80\% higher than that of Brand #2 Tirewear.
4. Examine the proc step

```
proc univariate data=mylib.fueldat cibasic normal mu=4 5;
var Income Roads;
title 'Use of Proc Univariate to Examine Distributions';
run;
```

Explain statistics produced by the proc univariate step above for each of the two variables given below:

(a) (3) Income

95% C.I.'s, t-test & p-value for \( H_0: \mu = 4 \) vs. \( H_a: \mu \neq 4 \), Test for normality

(b) (3) Roads

95% C.I.'s, t-test & p-value for \( H_0: \mu = 5 \) vs. \( H_a: \mu \neq 5 \), Test for normality

5. The graph below of side-by-side boxplots summarize primary highway miles of the 48 states according to the category variable per capita income level of the state.

(a) (4) Compare the shapes of the distributions for the three categories of level of income. Find the state with the highest number of primary highway miles and give an estimate of it.

All 3 shapes appear to be fairly symmetric, with one outlier in the middle income group & long right tail

(b) (4) Compare the location of the distributions for the three categories of level of income. Which category appears to have more states with higher number of primary highway miles? With lower number of primary highway miles?

Higher Income Group for both, higher & lower miles. Similar (blw. 4000-6000) for states in each income group.

(c) (4) Compare the spread (or dispersion) of the distributions for the three categories of level of income. Which category of income level (with respect to level) have more states with a large range of primary highway miles?

The spread (as measured by IQR) increases from low to high with the high income group having the largest range

(d) (4) Name 3 features in the boxplots below you could use to verify the model assumptions made when analyzing this data using ANOVA methods for comparing the three groups of states. Which of these appear to be the least plausible?

Normality, Spread, Outliers. The spreads appear to be the assumption least plausible.
6. Consider following data:

<table>
<thead>
<tr>
<th>x</th>
<th>26.5</th>
<th>18.3</th>
<th>21.8</th>
<th>19.6</th>
<th>19.4</th>
<th>30.7</th>
<th>20.2</th>
<th>16.8</th>
<th>16.9</th>
<th>14.0</th>
<th>15.4</th>
<th>15.2</th>
</tr>
</thead>
</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 \( \epsilon \) 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?

\[ R^2 = 0.4442 \] 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 \( \sigma^2 \)?

\[ SSE = 34.8 \quad 476.15 \quad \text{with 10 d.f.} \]

\[ \hat{\sigma}^2 = \hat{\sigma}^2 = 34.84762 \]

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.

\[ F = 7.99 \quad p \text{-value} = 0.0179 \]

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

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. \( H_a : \beta_1 \neq 0 \).

\[ t = \frac{\hat{\beta}_1}{\hat{\sigma}_1(\hat{\beta})} = \frac{-1.03259}{0.36524} = -2.8272 \]

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

\[ \hat{\beta}_1 \pm t_{0.025, 10} \times \hat{\sigma}_1(\hat{\beta}) \]

\[ -1.03259 \pm 2.12 \times 0.36524 \Rightarrow (-1.7996, -0.2656) \]

vi) (4) Use the value of \( h_3 \) to compute the standard error of the residual for observation 3.

\[ A \cdot \epsilon (e_3) = A \sqrt{(I - h_{33})} = \sqrt{34.84762 \cdot 1.1024} = 5.5928 \]

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

\[ A \cdot \epsilon (\hat{y}_3) = A \sqrt{h_{33}} = \sqrt{34.84762 \cdot 1.1024} = 1.8890 \]
viii) (4) Use the residual for observation 3 and its standard error to calculate the corresponding studentized residual.

\[ r_3 = \frac{e_3}{s.e.(e_3)} = \frac{-2.9239}{5.5928} = -0.5228 \]

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 (Rstudent in the y direction)
Both outside cut-off value of ±2
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 hats = \( \frac{2\hat{\sigma}^2}{n} = \frac{4}{12} = 0.3333 \)
Case #6 \( \hat{h}_{66} = 0.5578 \) exceeds 0.3333 so x-outliier
Plot C 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 = \( \frac{4}{n} = 0.3333 \)
Cook's D for Case #6 = 2.5 exceeds above it 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 as 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 seems 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 D, 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.