MATH 4044 – Statistics for Data Science

Practical Week 2 Solutions

Exercise 1
Statistics can be used to filter spam from incoming email messages. By noting specific characteristics of an email, a data scientist may be able to classify some emails as spam or not spam with high accuracy. One of those characteristics is whether the email contains no numbers, small numbers, or big numbers. Another characteristic is whether or not an email has any HTML content.

Data file for this exercise is based on a sample of 50 emails stored in a SAS data file called `email50.sas7bdat` located in `mydata` library on the SAS OnDemand server. The data statement to access this file is `data=mydata.email50`

Some of the variables in that file are as follows:

Variable	Description
spam	Specifies whether the message was spam; 0 = no, 1 = yes
num_char	The number of characters in the email
line_breaks	The number of line breaks in the email (not including text wrapping)
format	Indicates if the email contained special formatting, such as bolding, tables or links, which would indicate the message is in html format; 1 = html, 0 = text
number	Indicates whether the email contained no number, a small number (under one million) or a large number; none = no number, small = number under one million, big = large number

(a) Obtain a frequency distribution table of variables *spam* and *format*, with *spam* as the row variable. Which would be more helpful to someone hoping to classify email as spam or regular email: row or column percentages?

Frequency distribution table for spam and format

The FREQ Procedure

Table of spam by type

spam		type: text	type: html	Total
No	Frequency	9	36	45
	Percent	18.00	72.00	90.00
	Row Pct	20.00	80.00	
	Col Pct	69.23	97.30	
Yes	Frequency	4	1	5
	Percent	8.00	2.00	10.00
	Row Pct	80.00	20.00	
	Col Pct	30.77	2.70	
Total	Frequency	13	37	50
	Percent	26.00	74.00	100.00

Figure 1. Frequency distribution table for the *spam* and *type* variables

Such a person would be interested in how the proportion of spam changes within each email format. This corresponds to column percentages, which are based on the proportion of spam in plain text emails and the proportion of spam in HTML emails.

Examining column percentages, we see that a higher percentage of plain text emails were spam (30.77%) compared to HTML emails (2.70%).

This information on its own is insufficient to classify an email as spam or not spam, as nearly 70% of plain text emails are not spam.

(b) Obtain the clustered bar chart and 100% stacked bar chart of the same variables. Which would be more helpful?

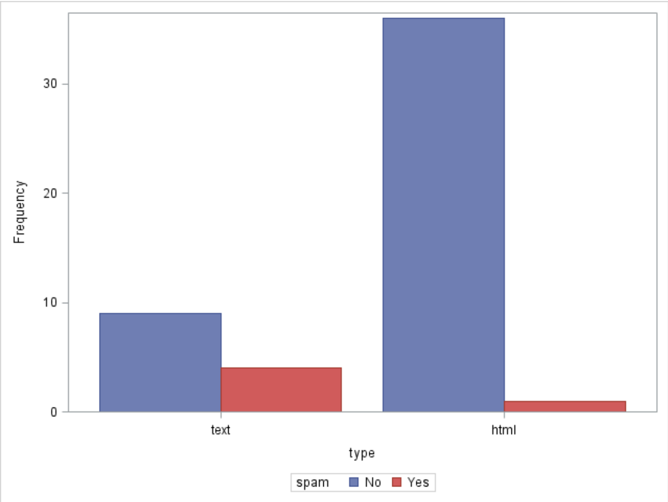


Figure 2. Clustered bar chart of the *spam* and *type* variables

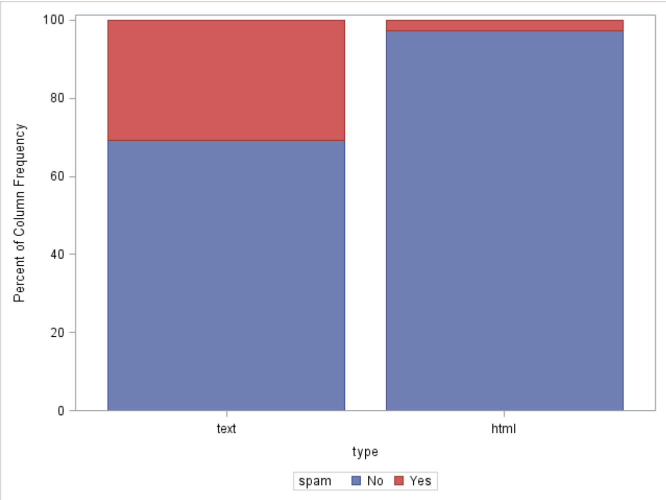


Figure 3. 100% stacked column chart of the *spam* and *type* variables

The 100% stacked column chart is more helpful as it allows comparisons between text and HTML format emails. It is immediately apparent that only a very small percentage of HTML emails were spam, whereas they were much more common among plain text emails.

The clustered bar chart is more difficult to interpret as it is based on frequencies or counts rather than percentages. As there were much fewer plain text emails (13 vs 37 for HTML format), it is not possible to make meaningful comparisons between the two email formats based on heights of the bars in the chart.

(c) Repeat parts (a) and (b) with variable *number* instead of *format*.

Frequency distribution table for spam and number

The FREQ Procedure

Table of spam by number					
		number			
		big	none	small	Total
spam					
No	Frequency	6	3	36	45
	Percent	12.00	6.00	72.00	90.00
	Row Pct	13.33	6.67	80.00	
	Col Pct	85.71	50.00	97.30	
Yes	Frequency	1	3	1	5
	Percent	2.00	6.00	2.00	10.00
	Row Pct	20.00	60.00	20.00	
	Col Pct	14.29	50.00	2.70	
Total	Frequency	7	6	37	50
	Percent	14.00	12.00	74.00	100.00

Figure 4. Frequency distribution table for the *spam* and *number* variables

We would be interested in how the proportion of spam changes within each number category. This again corresponds to column percentages, which are based on the proportion of spam in emails with no numbers, emails with small numbers and emails with big numbers.

Examining column percentages, we see that emails with small numbers were spam 2.7% of the time (relatively rare). We also see that 50% of emails with no numbers were spam, and 14.29% of emails with big numbers were spam.

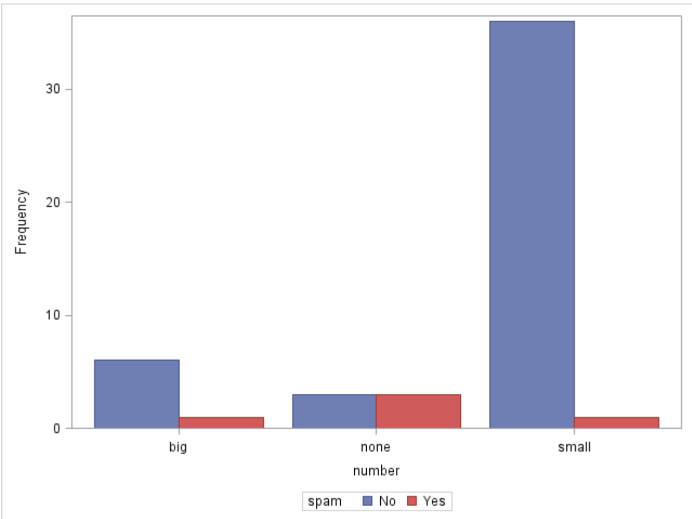


Figure 5. Clustered bar chart of the *spam* and *number* variables

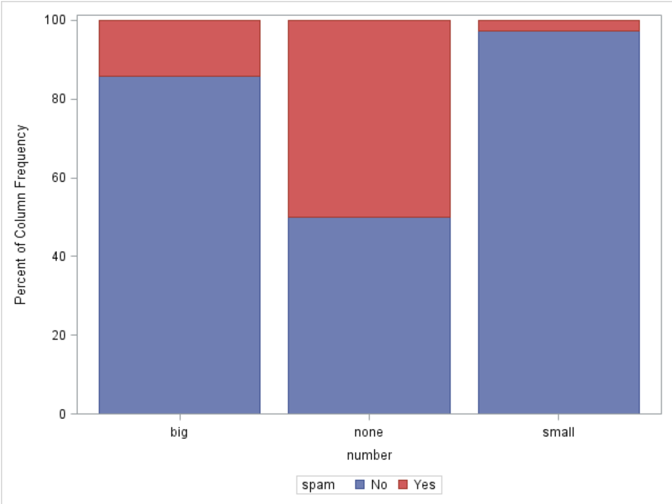


Figure 6. 100% stacked column chart of the *spam* and *number* variables

The 100% stacked column chart is again more helpful as it allows comparisons in relation to the inclusion of numbers in emails. It shows immediately that emails with small numbers are very rarely spam.

(d) Would either characteristic, *format* or *number*, alone be effective in identifying spam email? Explain briefly.

Neither characteristic alone is sufficient to identify an email as spam, although *number* alone may be more useful. If we consider format and number together (with many other variables), we stand a reasonable chance of being able to classify some email as spam or not spam. [There are statistical procedures that would allow us to do this.]

Exercise 2
Data file for this exercise is called `marathon.sas7bdat` and stored in `mydata` library. The data statement to access this file is `data=mydata.marathon`

It contains finishing times, in hours, for male and female winners of the New York marathon between 1980 and 1999.

(a) Obtain a histogram and boxplot of finishing times. What features of the distribution are apparent in the histogram and not in the boxplot? What features of the distribution are apparent in the boxplot and not in the histogram?

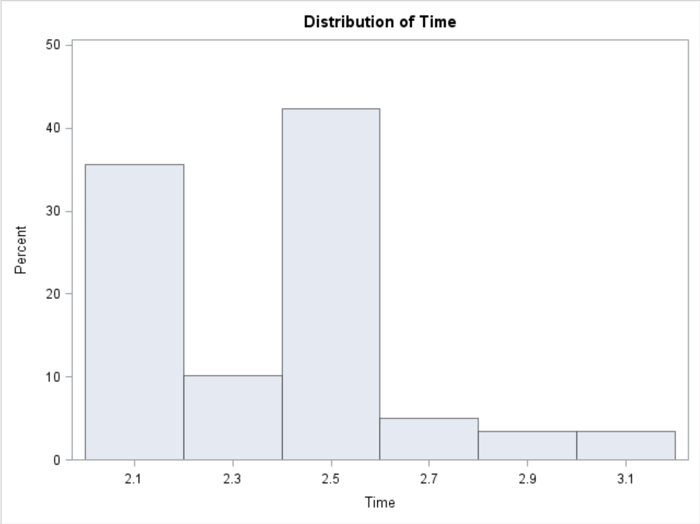


Figure 7. Histogram of finishing times (in hours)

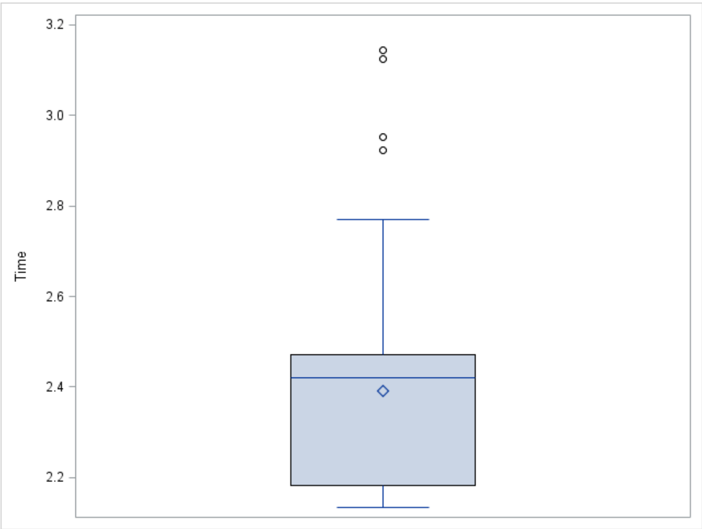


Figure 8. Boxplot of finishing times (in hours)

The histogram shows two distinct peaks in the distribution suggesting that marathon runners come from two distinct populations. The boxplot is not able to capture this characteristic of the distribution of finishing times.

The boxplot indicates that there are at least four outliers – runners who took much longer than the majority to finish the marathon – in the distribution of finishing times. The histogram shows a long right tail which suggests there could be outliers present, but we cannot be sure until we examine a boxplot.

(b) The distribution of finishing times is bimodal – it has two distinct peaks. What may be the reason for the bimodal distribution? Explain.

The data file includes finishing times of both male and female marathon runners. We would expect male marathon runners to be generally faster, which could explain two peaks in the distribution of finishing times.

(c) Obtain a boxplot of finishing times by gender and compare the distribution of marathon times for men and women. Comment briefly.

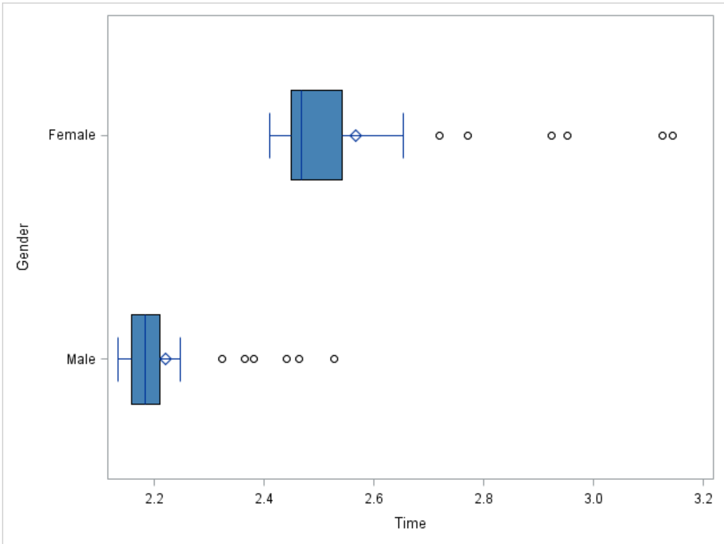


Figure 9. Boxplot of finishing times by gender

The distribution of finishing times for males is nearly symmetric but with a number of outliers, finishing times of over 2.3 hours. The distribution of finishing times for females is skewed to the right and has a number of outliers. Comparing the positions of the median lines and width of the boxes we see that finishing times for females are longer and have more dispersion.

[Note the position of the means. Both are larger than Q3 and their values have been inflated by the outliers. It would be difficult to argue that the means represent 'typical' finishing times. This demonstrates that the mean is not a 'robust' measure of centre.]

Exercise 3

Data file for this exercise is called `cars.sas7bdat` and comes from the `sashelp library`.

The data statement to access this file is `data=sashelp.cars`

Suppose we wish to investigate fuel economy of cars in city vs highway driving conditions based on their origin (Asia, Europe and US). Variables of interest are therefore *Origin*, *MPG_City* and *MPG_Highway*.

(a) Obtain Descriptive Statistics, histograms and boxplots of *MPG_City* by *Origin*. Use a variable of your choice to identify outliers.

The MEANS Procedure

Analysis Variable : MPG_City MPG (City)										
Origin	N Obs	Mean	Std Dev	Minimum	Maximum	N	N Miss	Lower Quartile	Median	Upper Quartile
Asia	158	22.013	6.733	13.000	60.000	158	0	18.000	20.500	24.000
Europe	123	18.732	3.290	12.000	38.000	123	0	17.000	19.000	20.000
USA	147	19.075	3.983	10.000	29.000	147	0	17.000	18.000	21.000

Figure 10. Descriptive statistics for variable *MPG_City*

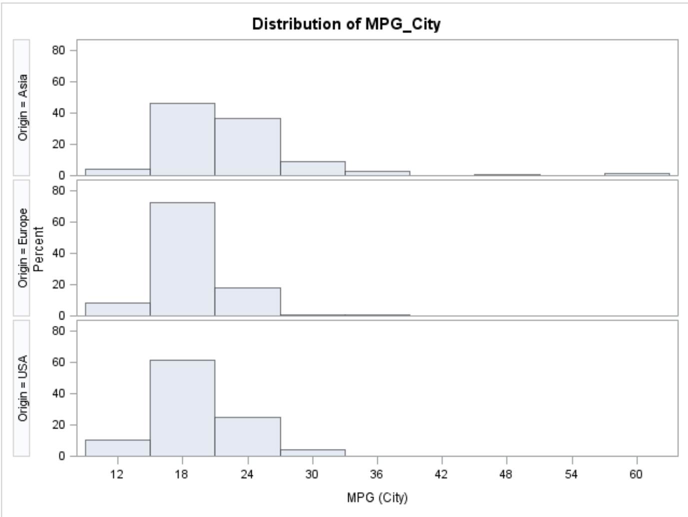


Figure 11. Distribution of *MPG_City* by *Origin*

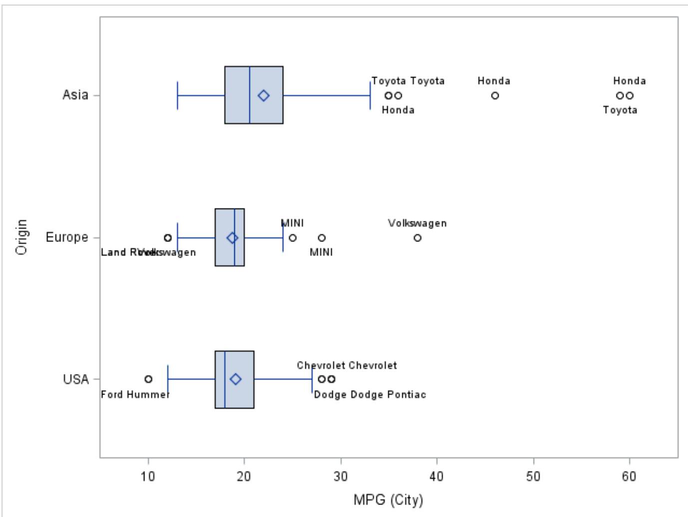


Figure 12. Boxplots of *MPG_City* by *Origin*

(b) Obtain Descriptive Statistics, histograms and boxplots of *MPG_Highway* by *Origin*. Use a variable of your choice to identify outliers.

The MEANS Procedure

Analysis Variable : MPG_Highway MPG (Highway)										
Origin	N Obs	Mean	Std Dev	Minimum	Maximum	N	N Miss	Lower Quartile	Median	Upper Quartile
Asia	158	28.266	6.771	17.000	66.000	158	0	25.000	27.000	31.000
Europe	123	26.008	4.168	14.000	46.000	123	0	24.000	26.000	29.000
USA	147	26.014	5.397	12.000	37.000	147	0	22.000	26.000	29.000

Figure 13. Descriptive statistics for variable *MPG_Highway*

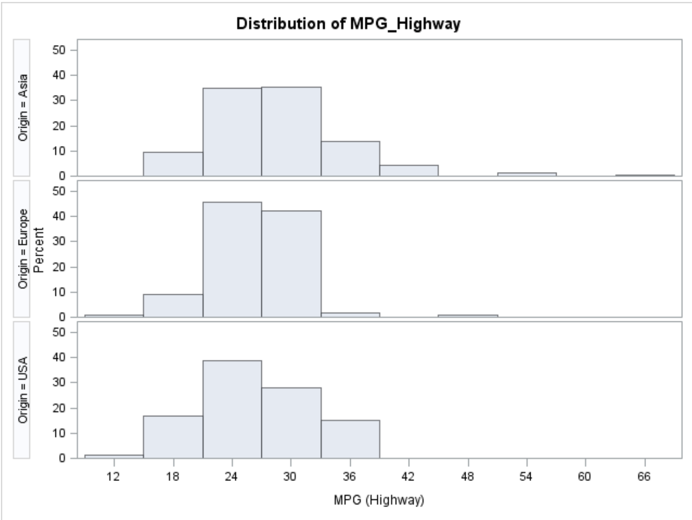


Figure 14. Distribution of *MPG_Highway* by *Origin*

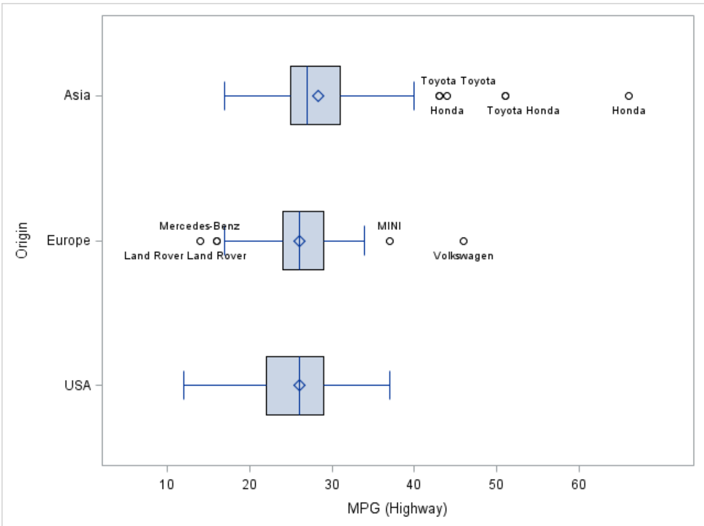


Figure 15. Boxplots of *MPG_Highway* by *Origin*

(c) Discuss your results from parts (a) and (b). What are some of your key observations?

Some observations (which statistical concepts am I using to make them?):

- Distributions of MPG in city driving conditions for cars made in Asia and the USA are skewed to the right while the distribution of cars made in Europe is nearly symmetric. All distributions have outliers.

- The distribution of MPG in city driving conditions for cars made in Asia appears to have the most dispersion; the distribution for cars made in Europe has the least.
- Cars made in Asia appear to be the best performers in city driving conditions. Various models of Honda and Toyota perform exceptionally well. Volkswagen is a stand-out for European cars.
- There are European and American makes that have very low fuel efficiency in city driving conditions (e.g. Mercedes or Ford).
- All cars, regardless of origin, generally do better on a highway. Typical highway MPG is approximately 26, compared to 18 to 20 in city driving conditions.
- The distribution of highway MPG for American cars has the most dispersion but no outliers, so no make is exceptionally good or bad in terms of fuel efficiency in highway driving conditions.
- Various makes of Honda and Toyota (Asian cars) and Volkswagen (European cars) again perform exceptionally well.

Appendix

Code for Exercise 1:

```
/* Data step to create a new data file with one variable renamed */

data work.temp_email;
/* New SAS data det to be created */

      set work.email50;
      /* Read observations from email50 data set */

      type=format;
      /* Rename variable format as type to avoid confusion with
      proc format */

      format spam SpamF. type TypeF.;
      /* A full stop '.' MUST follow each format name */

/* Associate new labels with variable values. New formats will
be permanently assigned to the variables in the new data file*/

run;

proc format;
/* Create formats or labels to be associated with values */

      value SpamF 0 = 'No' 1 = 'Yes';
/* Format name and new formats for values of spam; instead of 0 and 1 we
will now have No and Yes in our output */

      value TypeF 0 = 'text' 1 = 'html';
      /* Format name and new formats for values of type */

run;

/* Exercise 1 part a & b */

title 'Frequency distribution table for spam and format';

proc freq data=work.temp_email;     /* Use the new data set */
      tables spam * type / out=freq outpct;
run;

title '100% stacked bar chart of spam by email format';

proc sgplot data=freq;
/* Use frequencies stored at previous step */
      vbar type / response=pct_col group=spam;
      /* Use column frequencies */
run;

title 'Clustered bar chart';

proc sgplot data=work.temp_email;
      vbar type / group=spam groupdisplay=cluster;
run;

/* Exercise 1 part c */

title 'Frequency distribution table for spam and number';
```

```
proc freq data=work.temp_email;
      tables spam * number / out=freq outpct;
run;

title '100% stacked bar chart of spam by number';

proc sgplot data=freq;
      vbar number / response=pct_col group=spam;
run;

title 'Clustered bar chart';

proc sgplot data=work.temp_email;
      vbar number / group=spam groupdisplay=cluster;
run;
quit;
```

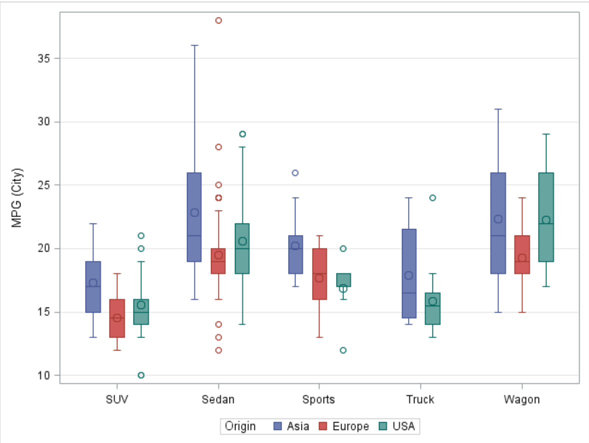
Code for Exercise 2:

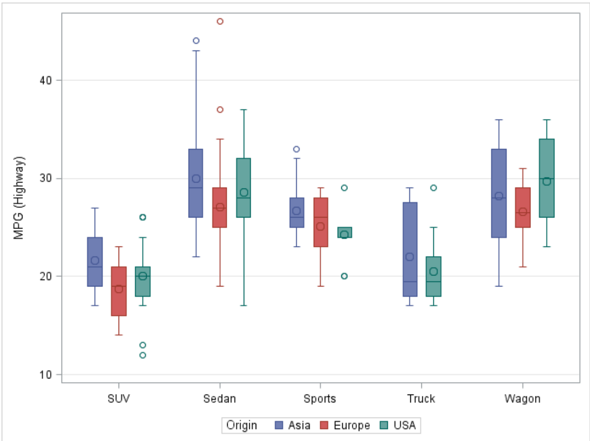
```
proc format;  /* Creating formats or labels for values */
      value $gender 'm' = 'Male' 'f' = 'Female'; /* For character variables
format names must start with $ */
run;

title 'Boxplot of finishing times by gender';

proc sgplot data=work.marathon;

      format gender $gender.;
      /* This statement is used to associate formats defined in $gender
with variable gender
         for the duration of the current procedure */

      hbox time / category=gender fillttrbs=fill (color=steelblue);

run;
```

Code for Exercise 3:

I have used ***Describe > Summary Statistics*** task for this exercise, and then modified the code generated by that task to make some adjustments.

Here are more 'fancy' boxplots based on the car data, showing the distribution of MPG by Origin and Type:





Here is the code that was used to generate these boxplots:

```
proc sgplot data=sashelp.cars (where=(type ne 'Hybrid'));
      vbox MPG_City / category=Type group=Origin grouporder=ascending;
      yaxis grid;
      xaxis display=(nolabel);
run;

proc sgplot data=sashelp.cars (where=(type ne 'Hybrid'));
      vbox MPG_Highway / category=Type group=Origin grouporder=ascending;
      yaxis grid;
      xaxis display=(nolabel);
run;
quit;
```