Lab 2: Describing Variables

We will explore in this lab the various ways that SPSS allows you to get information about your dataset's variables.

Here are some of the topics we will look at:

- Getting numerical summaries on the data
- Creating Histograms and Boxplots
- Customizing Graphs
- Copying Graphs for use in Word and PowerPoint

We will be using the cars dataset that you created last time. You should have saved it somewhere in native SPSS format. Make sure it is somewhere on the computer (not in your email and not in a thumb drive), and double-click it to open it up in SPSS.

This lab will assume that you have the data in front of you and that you have carried out all the steps described in Lab 1.

Getting Numerical Summaries

We already saw last week how to get a frequency table. This will we will use the same menu to compute various numerical summaries.

To compute numerical summaries:

- Go to "Analyze -> Descriptive Statistics -> Frequencies".
- Select the variable or variables you want to use. In our case, we will use "CMpG". Use the arrow to move it to the right side.
- Make sure the "Display frequency tables" box is NOT checked.
- In the "Statistics" submenu, select the statistics you want:
  - Mean and Median are over to the right
  - Std. Deviation and min/max are near the bottom
  - Quartiles are on the left. Make sure to at least select the "Quartiles" check-box.
  - If you want, you can add specific percentiles by checking that button and putting the values in, for instance "0.1" if you want the 10th percentile.
  - Click Continue.
- Click OK.

This should have created a new table for you. It should tell you:

- how many values there are (414),
- how many of them are missing values (14),

- that the mean is about 20.1 and the median is about 19 (which would be consistent with a skewed right distribution, think about this),
- the standard deviation is about 5.21,
- the quartiles are at 17 and 21, so IQR = 4,
- the minimum CMpG is 10, the maximum 60

So one conclusion we can draw for example is that 50% of the car models have mileage between 17 and 21.

Do a similar table for weights. What is a typical weight range? What is the 5 number summary?

Creating Histograms and Boxplots

For scalar variables, histograms are the standard way of visualizing them.

To do a **histogram**:

- Go to "Graphs -> Chart Builder"
- Select "Histogram" from the bottom left, then double-click on the first graph. This should add it to the main area.
- Drag the variable you want to work with to the x axis.
- Ignore the y axis. For 1-variable plots you should NEVER put anything in the y axis.
- Click OK, the graph should show up.

Look at the graph and the table you created earlier with the numerical summaries, and compare the two. Make sure that they match and make sense.

Describe the city mileage distribution in words (modes, skewness, outliers, typical range for most cars).

Do a similar graph for the weight. You can drop it on top of the other variable that was there before, it should just replace it. Does the graph for the weight match the table you made earlier? Any things that stand out in the graph?

To create a **boxplot**:

- Go to "Graphs -> Chart Builder"
- Select "Boplot" from the bottom left, then double-click on the first graph. This should add it to the main area.
- Drag the variable you want to work with (CMpG) to the y axis.
- Ignore the x axis for now.
- Switch to the tab that says "Groups/Point ID", and check the "Point ID label" box. This makes a new area appear in the graph, named "Point Label Variable?"
- Drag the CarName variable there. Because there might be outliers, they will be identified in the boxplot as circles with labels, this tells it what to use for the labels.

- Click OK, the graph should show up.
- If you want to draw one boxplot for each of the different categories of another variable, you use the x axis. So go back to the builder, and drag the Car type variable to the x axis.
- This should draw one boxplot per car type, each boxplot still showing the CMpG.

This is where boxplots really shine, when you want to compare say the CMpG across the different car types. In general you want to compare the values of a scalar variable against the different categories of a categorical variable. We will return to this theme later.

Can you tell from this which car types tend to have more efficient cars?

Now draw a similar boxplot for car weights in relation to the type. What conclusions can you draw from that?

Customizing Graphs

There is an almost infinite set of customizations you can do to graphs, we will just start scratching the surface.

To customize a graph:

- Double-click the graph to enter its "editing mode". Do this now for the boxplot showing CMpG and car type.
- One thing we can do is *change axis scales*. In this case let us try to make the y axis a bit smaller, so that it hides most of the outliers and starts closer to the minimum.
    - Click the line that is the y axis until you see it and only it selected; it should get a yellow boundary.
    - The properties window that has popped up should have a Scale section. Change the min and max values there to 5 and 35 respectively.
    - Hit "Apply". You should see the graph change and restict itself.
- All those labels are useful but a bit annoying. Right-click one of them and select "Hide Data Labels".
- The Median CMpG was 19, we would like to add a line there.
    - From the "Options" menu choose "y axis reference line".
    - In the Properties window, in the "Reference Line" tab, type in 19 or choose Median from the pulldown menu.
    - Click "Apply" to see the effects..
    - In the "Lines" tab of the Properties window you can choose a different color. Do so, and also make it dotted and thinner (weight 0.5).
    - Click "Apply" to see the effects.
- You can also change the color of the boxes, by clicking to select one of them.

- – The first time you click it selects all the boxes, you can change them all at once.
- – The second time you click it selects the specific box only. You can use this to give that particular box a different color.
- – Do NOT go color-crazy! Only use stylistic elements if they help make the point that you want your data to make. Elements that don't highlight the point are more of a distraction than anything else.

- We want to make the medians thicker. Click on one median, make sure they are all selected, and change the weight to something a bit bigger, say 2.5.
- We want to try to make the boxplots horizontal. There are two buttons near the top right of the Chart Editor that allow you to switch back and forth between horizontal and vertical. Try them out.

  - – The human eye is some times more comfortable comparing horizontal positions. Again, choose what makes your point more clear.

- You can rearrange the groups if you want to, or even remove some groups:

  - – Select a box, make sure they are all selected.
  - – In the Properties window, choose "Categories".
  - – Select a category in the Order box, and use the arrows to move it up or down.
  - – Use the X to exclude a category.

When you are happy with how your graph looks, you can exit the Chart editor by just using the X button at its top right.

Copying Graphs

One really important thing to remember is this:

- Graphs do not resize well once they are outside of SPSS.
- It is really important to keep the graphs as they should be in SPSS, and save the output viewer document in addition to the data.
- Resize your graphs in SPSS first, to the size you want to use, before copying them over. You might have to do this twice, once for Word and once for PowerPoint.
- To copy the graph, simply select it by clicking it, then right-click and select "Copy". Then just do a Paste into Word or PowerPoint or wherever.
- For PowerPoint you might end up with small fonts. You will want to make it even smaller in SPSS, then stretch it out in PowerPoint.
- Alternatively you can edit the graph, and select the two axis labels one at a time, and change their "preferred size".