## ggplot part 2 - Solutions

Make sure ggplot2 library is installed and then loaded correctly!

Hopefully you have saved the data and RMD into the same folder, and created a project file for that folder. If so then the BeanSurvey dataset can be read in without modifying the code below.

library(ggplot2)  
BeanSurvey<-read.csv("bean\_survey.csv")

**Question 1: I am trying to make a histogram of the farmer’s ages with different panels for each village. Can you identify and fix the errors in my code?**

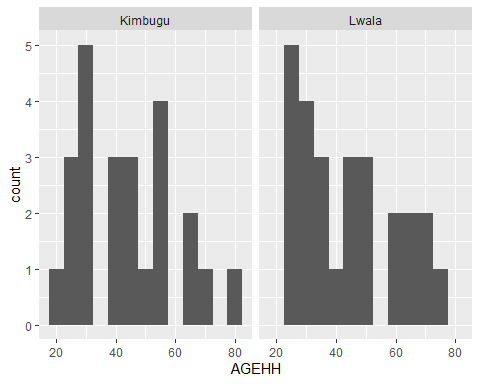
ggplot(BeanSurvey,aes(x=AGEHH))+ geom\_hist(binwidth =5)+ facet\_wrap(VILLAGE)

\*\* Answer

Remember to try to find the errors by running the code and checking the messages. It can be very difficult to just magically spot the error by starting at the code without some help from the error message!

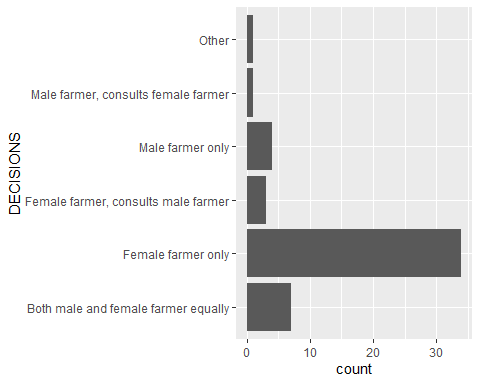
There were two mistakes: Firstly the correct geometry should be geom\_histogram() rather than geom\_hist(). The autocomplete within the code window could help you with this. Secondly the tilda ~ was missing from within facet\_wrap() - remember this is needed in facet\_wrap before the variable name.

ggplot(BeanSurvey,aes(x=AGEHH))+  
 geom\_histogram(binwidth =5)+  
 facet\_wrap(~VILLAGE)



**Question 2a: I am making a barplot of decision making, by gender of the household head. Can you modify the code, so that it looks like the example in the workbook with female headed households coloured in ‘purple’ and male headed households coloured in ‘orange’**

ggplot(BeanSurvey,aes(y=DECISIONS))+  
 geom\_bar()

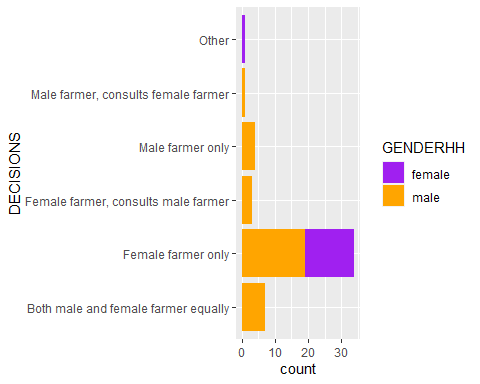


\*\* Answer

To get the required plot we need to make two modifications. Firstly we need to set the fill aesthetic to be based on the GENDERHH column from the data. The bars have two sorts of colour we can modify - fill for the inner region and colour for the outline. If we changed colour this would still have the bars shaded in grey but with red and blue outlines. This is not what we want! We want the shading colour, the fill to be modified. And because we are varying the colour based on a column from the data we set it in the aesethetics not in the geometry.

To use the purple and orange colours we need to modify the fill aesthetic using the correct scale\_ function. Because we are chaning the fill aesthetic manually then the function we need to add is scale\_fill\_manual. Inside this function we have two colours that we need specify to the values argument because there are two values for GENDERHH, so we need to use the c() function to bring these colours together.

ggplot(BeanSurvey,aes(y=DECISIONS,fill=GENDERHH))+  
 geom\_bar()+  
 scale\_fill\_manual(values=c("purple","orange"))



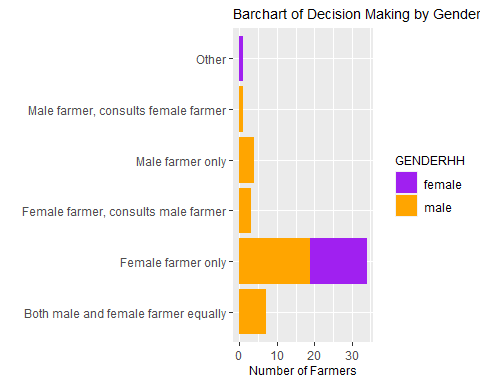
**Question 2b: Taking the plot you created in Question 2a, now make some changes to the labels: i) remove the y axis label; ii) Change the x axis label to read “Number of Farmers”; iii) Add an informative title**

\*\* Answer

I can modify the y axis label so that it is blank "" rather than “remove” it - this is an easier way to achieve the same thing! So all three of my steps can be done in the same function - labs() and calling the y, x and title labels.

Depending on my screen resolution i might encounter some problems with my title running off the page. This is where it could be useful to also add a call to the theme function and make the text size smaller using element\_text(). An alternative solution could be to split the title into two components - a title and a subtitle.

ggplot(BeanSurvey,aes(y=DECISIONS,fill=GENDERHH))+  
 geom\_bar() +  
 scale\_fill\_manual(values = c("purple","orange"))+  
 labs(y="", x="Number of Farmers",title="Barchart of Decision Making by Gender of Head") +  
 theme(title = element\_text(size=9))



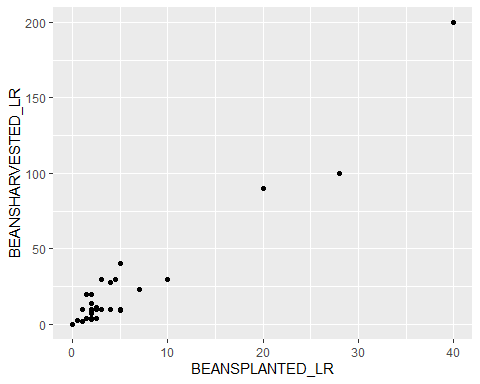
**Question 3: Make a scatter plot of the harvested quantities of beans in the long rains against the planted quantity of beans in the long rains. Place the harvested quantity on the y axis, and the planted quantity on the x axis.**

\*\* Answer

Here I need to identify the correct columns from the data set and then use the point geometry to make the scatter plot

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=BEANSPLANTED\_LR)) +  
 geom\_point()

## Warning: Removed 3 rows containing missing values (geom\_point).



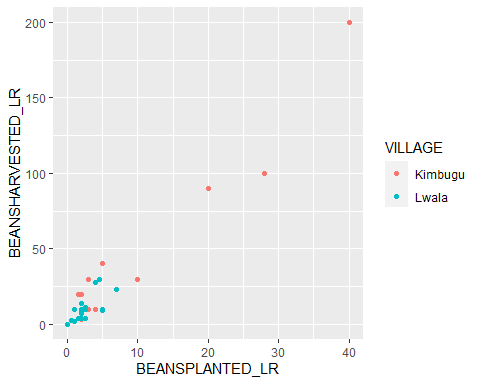
**Now consider how to also show how the different villages are associated with this this relationship. Try two different options: a) Make different coloured points for each village**

\*\* Answer

The key thing here is to understand that for the default plotting symbol of geom\_point the way to modify the colour is through colour rather than fill. So differently to the barchart from Q2, this time I need to map the colour aesthetic to the VILLAGE column.

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=BEANSPLANTED\_LR,colour=VILLAGE)) +  
 geom\_point()

## Warning: Removed 3 rows containing missing values (geom\_point).

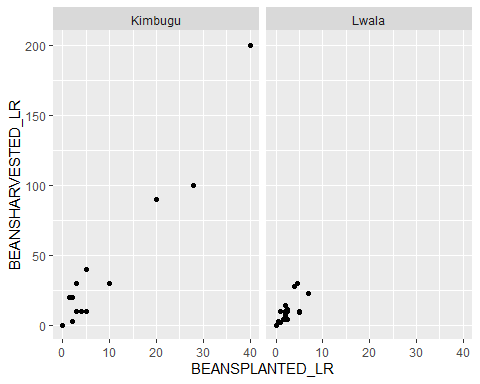


**b) Put the two villages in different panels**

\*\* Answer

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=BEANSPLANTED\_LR)) +  
 geom\_point() +  
 facet\_wrap(~VILLAGE)

## Warning: Removed 3 rows containing missing values (geom\_point).



Remember to use the tilde ~ before the VILLAGE when making the facet

**Consider which of these two plots shows the relationship more clearly. Take your preferred plot, and tidy up the axis labels and titles.**

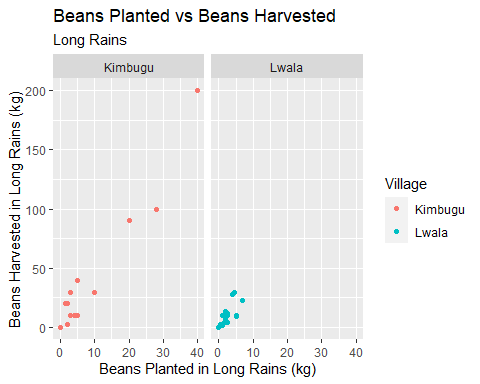
\*\* Answer

There are pros and cons to both plots (and lots of further ways to modify the plots as I talked about during the webinar). The ‘right’ plot here depends one exactly what we want to highlight - the facets show the within village patterns much more clearly, but make it harder to compare between villages. The colours allow you to compare the villages more easily, but make seeing the results within each village more difficult.

We could of course do both! Which is what I’ve done here and then used labs to label my axes. But of course this is more or less the same presentation as just using the facets, the colour is mostly just a bonus visual element.

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=BEANSPLANTED\_LR,colour=VILLAGE)) +  
 geom\_point() +  
 facet\_wrap(~VILLAGE)+  
 labs(x="Beans Planted in Long Rains (kg)",y="Beans Harvested in Long Rains (kg)",  
 title="Beans Planted vs Beans Harvested",subtitle="Long Rains",colour="Village")

## Warning: Removed 3 rows containing missing values (geom\_point).



**Question 4a: Make a plot showing how the quantity of beans harvested in the long rains is related to the household type (HHTYPE). Choose a sensible geometry to show this relationship.**

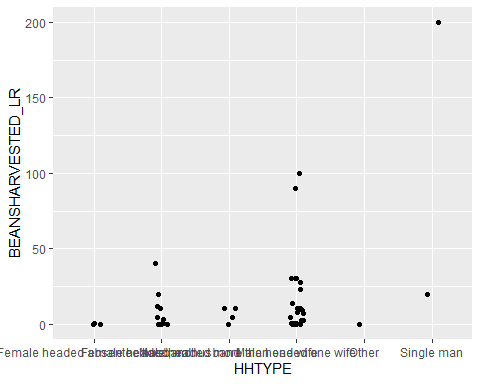
\*\* Answer

When considering what sort of plot to make we should always consider what type of variable we have. In this case we have a continuous numeric variable BEANSHARVESTED\_LR and a categorical variable HHTYPE. A common plot to make here might be a boxplot. However - this would actually be very misleading because of the way the data is distributed. There are very few observations of households in certain ‘types’ - just two single men and just one ‘other’. Because the boxplots provide a summary of the values in each group then we need to have enough observations so that we can form a reasonable and meaningful summary. When you have only a few observations per group probably the better option would be to use a point geometry.

In this case, particularly with the zero values, there are many points overlapping, like in the final question from the previous module. So a jittered scatter plot may be better using geom\_jitter. A better plot might be possible by combining or removing the small categories. We will learn about this next week!

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=HHTYPE))+  
geom\_jitter(width=0.1)

## Warning: Removed 3 rows containing missing values (geom\_point).

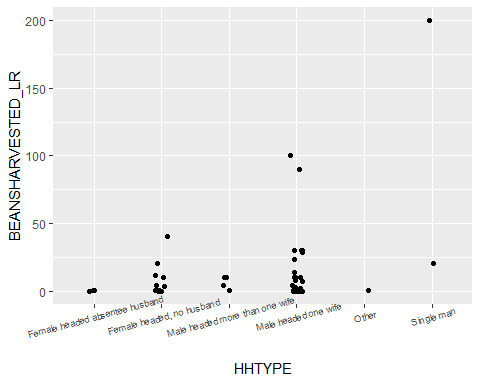


**Question 4b: Whatever plot you decided on for question 4a, you may have found that because the labels for the HHTYPE variable are quite long, the text along the axis became squashed. Look into the elements which can be customised within theme and try to make these labels fit better by decreasing the font size or modifying the angle at which the labels are aligned to the axis**

\*\* Answer

ggplot(BeanSurvey , aes(y=BEANSHARVESTED\_LR,x=HHTYPE))+  
geom\_jitter(width=0.1)+  
 theme(axis.text.x = element\_text(size=7,angle=15))

## Warning: Removed 3 rows containing missing values (geom\_point).



In the theme function we can find the component we want to modify, the axis text on the x axis, and then use element\_text to make this more legible. Using angle we can set this to be at 90 degrees (vertical) or at a slight angle, 10 to 15 degrees, and these could provide ways of making the labels easier to read. Just reducing the size of labels will probably make them too small to read!

**Question 5: Make a plot showing the relationship between the gender of the household head (GENDERHH), village, (VILLAGE) and whether the household sells any of their beans (SELLBEANS). Include nice colours, sensible axis labels, a title, and use one of the custom built-in themes to make the plot look nice. Also consider increasing the font size for some of the labels so they can be read more clearly.**

\*\* Answer

There are lots of ways you could have chosen to make a plot here! The way I went about it is just one option.

With these three variables the question I am probably most interested in is looking at whether gender of household head and village have any influence on the likelihood that the household sells beans. I have three numeric variables so I will be making some bar charts. I can have one variable as the bars, one on the x axis and the third as a facet. The bars should probably represent my outcome- “selling beans”. Structurally it would make most sense to have facets for village and gender on the x axis. The male and female headed households live within a village so grouping in this way is more logical and easier to understand than grouping panels by gender and having village on x axis.

I would quite like my bars to represent percentages rather than frequencies, because I have an unequal number of male and female headed households and i would like to compare them. Using “position=fill” within geom bar converts the bars from counts into proportions.

The axis labels are for a proportion - using scale\_y\_continuous i can change this into a percentage by setting the breaks and labels. This will also let me change from the breaks every 25% which is harder to read off the values as comapred to setting the breaks every 20%.

I can do a further trick when assigning colours - we are most interest in the % selling beans. And we can always work out the % not selling beans because it will add up to 100%. So when I set my colours - if i set a missing or fake colour for the “not selling beans group” the plot becomes a bit easier to read as I just see the bars making up the % selling beans.

I’ve finished off by using the light theme which i think looks quite nice, and adding in some labels.

ggplot(BeanSurvey,aes(fill=SELLBEANS,x=GENDERHH))+  
 geom\_bar(position="fill",show.legend=FALSE) +  
 scale\_y\_continuous(breaks=c(0,0.2,0.4,0.6,0.8,1),labels=c("0%","20%","40%","60%","80%","100%"))+  
 scale\_fill\_manual(values=c(NA,"darkred"))+  
 facet\_wrap(~VILLAGE)+  
 labs(y="% Selling Beans",title="Barplot of Bean Sales by Gender and Village",x="Gender")+  
 theme\_light()

