**Analyzing Heart Disease Dataset Answer Sheet**

# **Activity #1: Outline your analysis**

In this activity, students will need to write down the steps to check if there is an association between serum cholesterol level and resting blood pressure. This helps the student to break down the problem into its essential steps and think about how to solve it.

1. ✏️ **Question**: Think about the task of studying if there is an association **between serum cholesterol level and resting blood pressure**. Using the knowledge from previous lessons, break down the task. Write down the steps you will take.

**Sample answer:**

1. Create a Scatter Plot showing the relationship between serum cholesterol level and resting blood pressure.
2. Find the correlation coefficient between serum cholesterol level and resting blood pressure. Analyze the correlation coefficient to check if the relationship is strong or weak. Also, check the correlation coefficient to see if the relationship is positive or negative.
3. Perform a correlation test between serum cholesterol level and resting blood pressure. Review the correlation test (p-value) to see if there is enough evidence to conclude there is a significant linear relationship between serum cholesterol level and resting blood pressure.

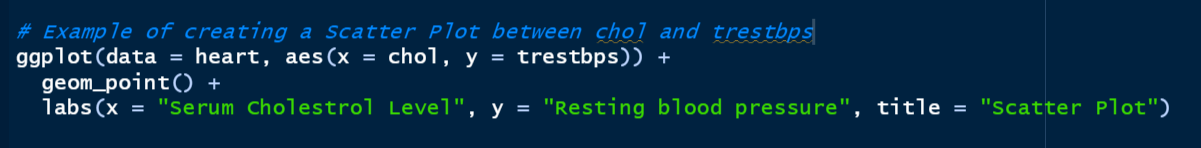
# **Activity #2: Use R to Perform Your Analysis**

Students should perform their own analysis based on the outline they created in Activity #1.

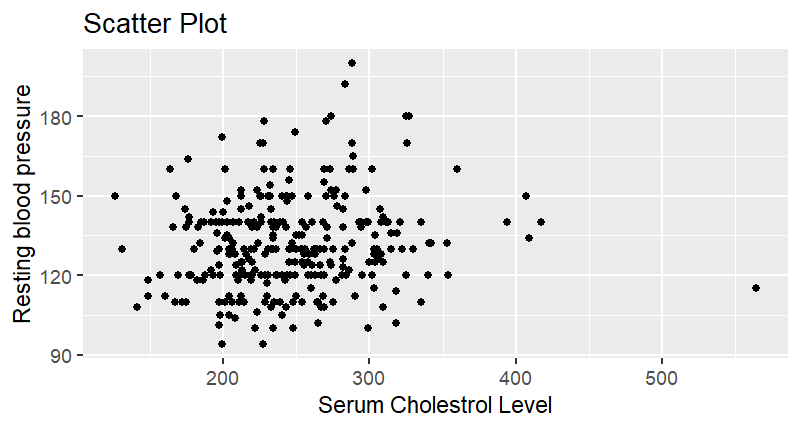
1. ✔️ Save and load the Heart Disease dataset provided by your teacher into Posit Cloud.
2. ✔️ Plot serum cholesterol level and resting blood pressure.
3. ✏️ Write down your observations.

**Sample Answer:**

This is the code to plot a scatter plot between serum cholesterol level and resting blood pressure:



This is what the scatter plot should look like after running the code above:

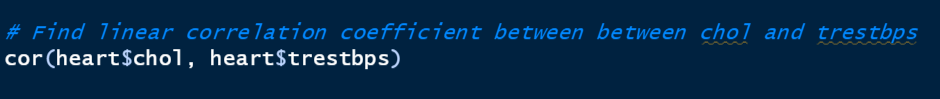


The box plot does not suggest a strong relationship or pattern between serum cholesterol level and resting blood pressure.

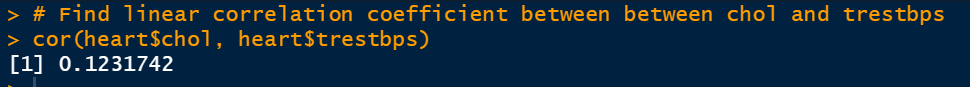
1. ✔️ Based on what you saw in step 2, test to see if there exists a meaningful association between serum cholesterol level and resting blood pressure. Use the techniques you learned in prior lessons.
2. ✏️ Write down your observations.

**Sample Answer:**

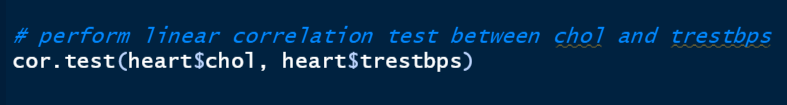
This is the code to find the correlation coefficient between serum cholesterol level and resting blood pressure:



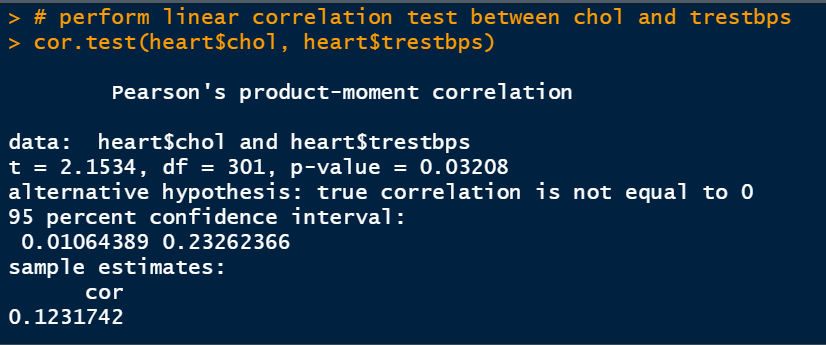
This is what the bottom left console panel should look like after running the code above:



From the correlation coefficient calculation, the correlation coefficient between serum cholesterol level and resting blood pressure is 0.1231742. This indicates a weak, positive correlation between these two variables. As serum cholesterol levels increase, resting blood pressure tends to increase slightly.

This is the code to perform a correlation test between serum cholesterol level and resting blood pressure:

This is what the console panel should look like after running the code above:



The correlation test between serum cholesterol level and resting blood pressure gives a p-value of 0.03208 and a correlation coefficient of 0.123. The 95% confidence interval extends from 0.01 to .23, meaning we are 95% confident that the correlation coefficient is somewhere in this range.

The p-value of 0.03208 is less than the commonly chosen significance level of 0.05 (or 5%). This suggests that there is evidence to conclude that there is a statistically significant positive linear correlation between serum cholesterol level and resting blood pressure.

# **Activity #3: Analyze and interpret your tests**

1. ✏️ Analyze the results of the test(s) you performed. What did you conclude? Justify your conclusion(s).

**Sample Answer:**

The results of the correlation test between serum cholesterol level and resting blood pressure indicate a statistically significant yet relatively weak positive relationship between these two variables. The positive correlation coefficient of 0.1231742 suggests that as serum cholesterol levels increase, resting blood pressure tends to increase slightly, and vice versa. This correlation is not likely due to random chance, as evidenced by the p-value of 0.03208, which is lower than the typical significance level of 0.05.

# **Activity #4: Outline your analysis**

In this activity, students will need to write down the steps to check if there is a relationship between serum cholesterol level and heart disease risk. This helps the student to break down the problem to its bare essentials and think about how to solve it.

1. ✏️ **Question**: Think about the task of studying if there is a relationship **between serum cholesterol level and heart disease risk**. Using the knowledge from previous lessons, break down the task. Write down the steps you will take.   
   Hint: Consider the differences between those who have a high risk of heart attack and those who have a low risk.

**Sample answer:**

1. Create a new column in the dataset that is a categorical variable that states if the patient has a “high” or “low” chance of a heart attack.
2. Create a Box Plot showing the variation of serum cholesterol levels for those with “high” and “low” chance of heart attack.
3. Perform a 2-sample t-test on serum cholesterol level between patients with a “high” chance of heart attack and a “low” chance of heart attack. Analyze the 2-sample t-test to see if the patients with a “high” chance of heart attack have a statistically different serum cholesterol level from patients with a “low” chance of heart attack.

# **Activity #5: Use R to Perform Your Analysis**

Students should perform their own analysis based on the outline they created in Activity #4.

**Question:**

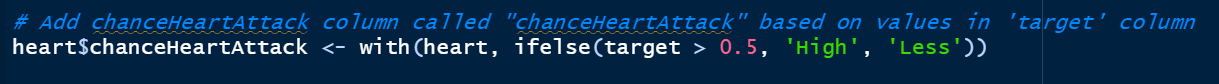
1. ✔️ Plot and compare the serum cholesterol levels of those with heart disease risk and those without.

* Hint: You may need to create a new categorical variable based on the numerical “target” variable. If you are having trouble doing this, ask the teacher for help

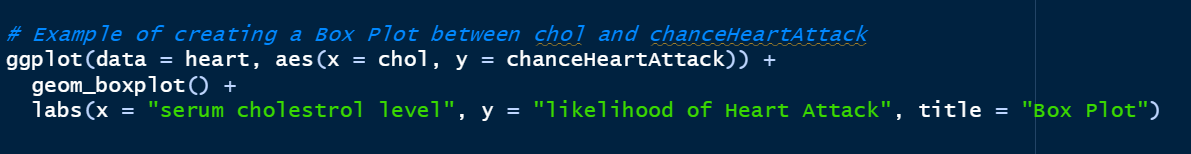
1. ✏️ Write down your observations.

**Sample Answer:**

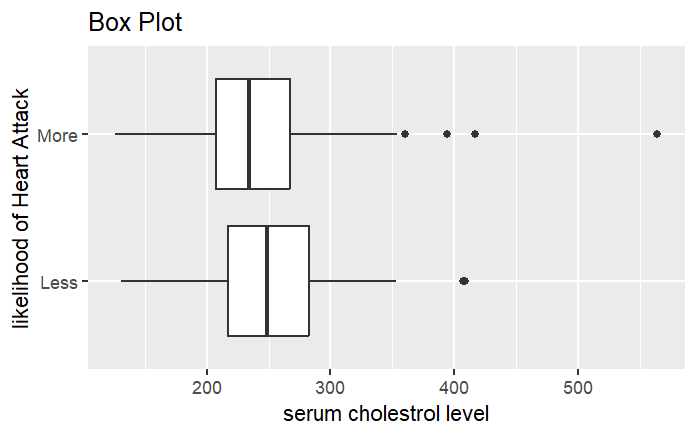
This is the code to create the new column:



This is the code to plot a Box Plot for the serum cholesterol level of patients with a “high” chance of heart attack and a “low” chance of heart attack:



This is what the box plot should look like after running the code above:



The box plot does not show much difference in serum cholesterol levels between patients with a “high” chance of a heart attack and patients with a “low” chance of a heart attack. The plot shows 5 outliers.

1. ✔️ Based on what you saw in step 1, test to see if there exists a meaningful difference in serum cholesterol levels between those with heart disease risk and those without. Use the techniques you learned in prior lessons.

* State your null and alternative hypothesis before you start.

1. ✏️ Write down your observations.

**Sample Answer:**

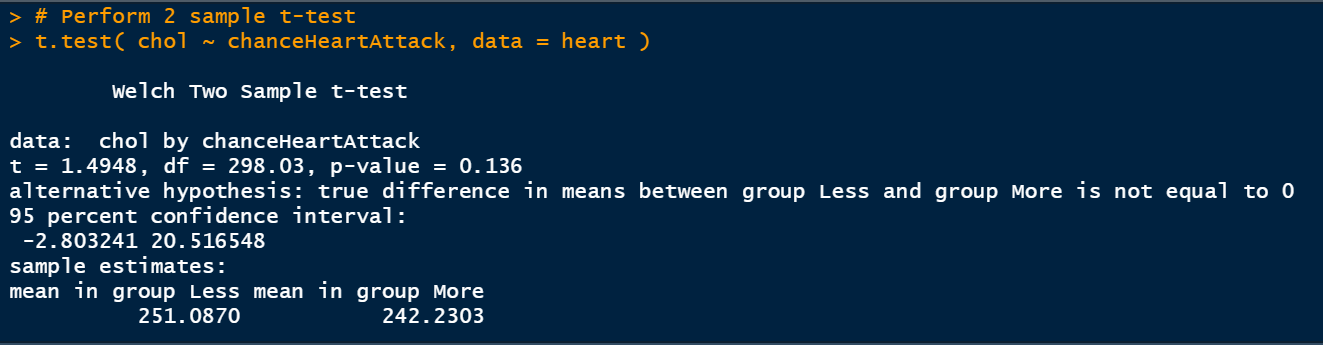
Null Hypothesis: There is no significant difference between the means of serum cholesterol level between patients with a “high” chance of heart attack and a “low” chance of heart attack.

Alternative hypothesis: There is a significant difference between the means of serum cholesterol level between patients with a “high” chance of heart attack and a “low” chance of heart attack.

This is the code to perform a 2-sample t-test on serum cholesterol levels between patients with a “high” chance of heart attack and a “low” chance of heart attack:



This is what the console panel should look like after running the code above:



A positive t-value of 1.4948 shows that, on average, the group with high heart disease risk has a higher serum cholesterol level compared to the group with low heart disease risk. However, the magnitude of this difference is not very large.

The p-value of 0.136 is greater than the commonly chosen significance level of 0.05 (or 5%). This suggests that there isn’t significant evidence to reject the null hypothesis (no difference in cholesterol levels between the two groups). Thus, we accept the null hypothesis - There is no significant difference between the means of serum cholesterol level between patients with a “high” chance of heart attack and a “low” chance of heart attack.

1. ✔️ Redo Activities #4-5 but use other variables that may be related to heart disease risk.

Students should do a similar analysis in the steps above.

# **Activity #6: Analyze and interpret your tests**

1. ✏️ Analyze the results of the test(s) you performed. What did you conclude? Justify your conclusion(s).

**Sample Answer:**

Based on the results of the 2-sample t-test between serum cholesterol level and heart disease risk, there is no significant difference in the mean serum cholesterol levels between individuals with high heart disease risk and those with low heart disease risk. The positive t-statistic (1.4948) suggests that, on average, the group with heart disease risk tends to have higher serum cholesterol levels. However, the p-value of 0.136 is greater than the commonly chosen significance level of 0.05, indicating that this observed difference in means is not statistically significant at the 5% level. Therefore, the results do not provide strong enough evidence to definitively conclude a significant difference in serum cholesterol levels between the two groups.

1. ✏️ How do you think the knowledge and skills gained from this lesson can be applied to real-world scenarios in the fields of medicine, research, or data analysis?
2. ✏️ Overall, how has this activity helped you develop a deeper appreciation for the importance of analyzing and interpreting data in the context of heart disease research? What other areas do you think data analysis could be valuable in the field of biology and medicine?