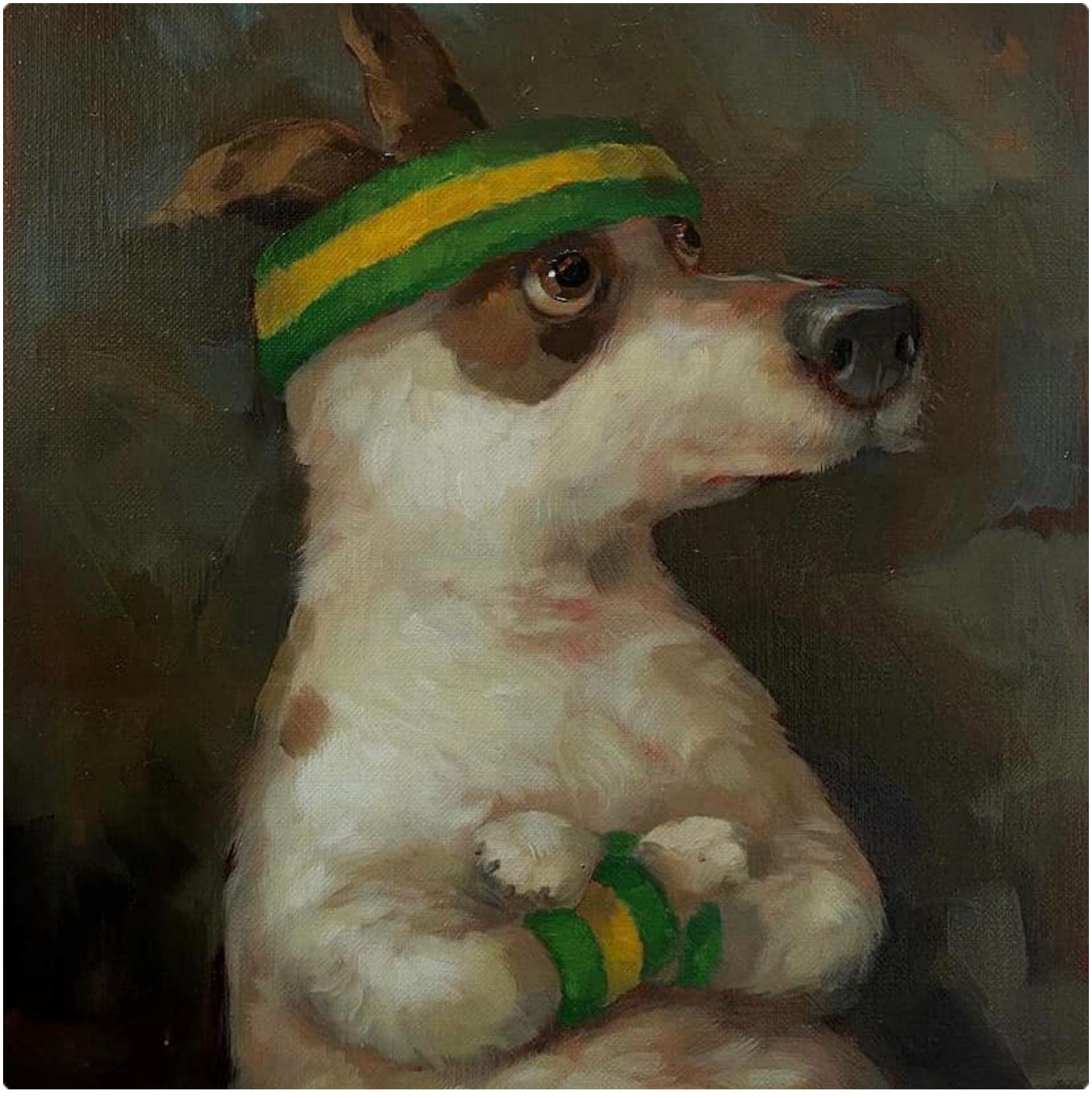
Lab 6 | Data Science

Statistical Exploratory Data Analysis



Main Goal & Objectives

The goal in this lab is to use hypothesis testing to conduct a exploratory data analysis. For this, we would be considering life expectancy, but this time we work with a new data set of US county level information.

Q1 | When we constructed the above_average_life-expectancy indicator, was it appropriate to use the mean for this purpose?

Yes, that would be the case. Whenever we use terms such as "average", this often indicates that we're taking the mean of the dataset that we're looking at. When we constructed the above_average_life-expectancy indicator, we wanted to look at the average life expectancy of all counties combined in the United States, and then compare the computed total average against the life expectancy in each county to see if it was below average or above average. As a result, it would be fair game to assume it is appropriate.

Q2 | When we constructed the above_average_life-expectancy indicator, how could you have used hypothesis testing to build this variable? If you had, how many categories would it have?

With regards to using hypothesis testing, we could have modeled the situation as follows:

- **Null Hypothesis:** the life expectancy of the county is equal to the national average
- **Alternative Hypothesis:** the life expectancy of the county is below the national average

The type of hypothesis testing we could use is one sample t-test, because we're comparing the national mean of life expectancy against the life expectancy of different counties.

Using the hypothesis testing we have, this would mean we have two categories of counties. The first category is the counties which have a life expectancy equal to the national average, and the other is the life expectancy of the county below the national average.

Q3 | What are we implicitly assuming by filling the missing values with 0 when we create the largest industry variable?

To fill missing values with zero means that it will not cause significant bias when it comes to calculating the mean, variance or standard deviation of that certain sample or dataset. A missing value also might implicitly imply that there are zero workers in that certain industry.

Q4 | Explain what it means to reject the null hypothesis in a Pearson's Correlation Coefficient Hypothesis Test

To reject the null hypothesis in a Pearson's Correlation Coefficient Hypothesis Test means that there is evidence to claim that there is a significant linear relationship between the two variables being tested. Furthermore, it also means that there is a statistically significant result.

The Null & Alternative hypothesis is as follows:

- **Null Hypothesis:** there is no linear relationship between the two variables
- **Alternative Hypothesis:** there is a significant linear relationship between the two variables

Q5 | Discuss the meaning of Pearsons's Correlation Coefficient Values obtained in your analysis. How does it help in interpreting the relationship between the variables under study?

Pearson's Correlation ranges from -1 to 1, where:

- When r = 1, this means there is a a positive linear relationship, which would imply that as one variable increases, the other variable also increases.

- When r = -1, this means there is a negative linear relationship, which would imply that as one variable increases, the other variable would decrease.

- When r = 0, this means there is no linear relationship between the variables. Changes in one variable do not systematically correlate with changes in the other.

Using the values of r helps to determine the type of relationship between two variables, and how one changes depending on the other. It also tells us how strongly related the variables are, and if they have a linear relationship.

Q6 | For which variables did the obtained correlation coefficient align with your initial expectation? Why or why not?

Some of my initial expectations were that life expectancy would have a positive linear relationship between Smokers, Adults with Obesity & Physically Inactivity. The correlation coefficient for those three were -0.737735, -0.482122 & -0.623696. These values make sense, as smoking, obesity and physical inactivity often have negative impacts on health and lead to a decreased life expectancy. Other such instances are drives alone and the average number of mentally unhealthy days.

My other initial expectations were that life expectancy would have a strong positive linear relationship with vaccination, however that was not the case. The correlation coefficient for that was 0.144385, as this value indicates that there is a weak linear linear relationship. I had this expectation because having a vaccination would help you live longer and protect you from diseases.

Q7 | Reflect on the assumptions of Pearson's Correlation vs Spearman's Correlation. How would you verify these assumptions in your analysis?

Assumptions of Pearson's Correlation is that the relationship between the two variables are linear and that both variables should be normally distributed. A linear relationship is when change in one variable should correspond linearly with change in the other variable.

Assumptions of Spearman's correlation is that the relationship between the two variables is monotonic. A monotonic relationship means that as one variable increases, the other variable increases or decreases. It doesn't necessarily have to be a linear manner.

To verify such assumptions in our analysis using a scatter plot, we can do this in the bullet points below:

- **Pearson's Correlation:** if the scatter plot shows a straight-line relationship, this means there is a linear relationship

- **Spearman's Correlation:** if the scatter plot shows a relationship where the two variables are increasing or decreasing, even if it's not a straight line

Q8 | Explain what it means to reject the Null Hypothesis in a Kruskal-Wallis Hypothesis Test

To reject the null hypothesis in a Kruskal-Wallis hypothesis test means that there is statistical evidence that the distributions or medians of groups that being compared are different in some way.

It is a indication that at least one of the groups differs significantly from the others in terms of median values. The null and alternative hypothesis for a Kruskal-Wallis test is as follows:

- **Null Hypothesis:** there is no difference between the groups and that the medians of all groups are equal

- **Alternative Hypothesis:** there is at least one group that is different from others, which means that the distributions or medians aren't the same

Q9 | Explain why the Kruskal-Wallis test was chosen as an appropriate non-parametric alternative for comparing more than two independent groups

Kruskal-Wallis is chosen as a hypothesis test because it doesn't require the assumption that the data is normally distributed or that there are equal variances.

Q10 | Regarding a p-value, fixing a threshold of 0.05 means that we are setting the risk of rejecting the null hypothesis when we shouldn't at 5%. In this section, you've conducted over 20 hypothesis tests. What may this imply about the likelihood you've made at least one type I error?

A type I error is when we reject the null hypothesis when it is true. When the p-value is less than 0.05, we reject the null hypothesis, and if the p-value is greater than 0.05, then we fail to reject to null hypothesis. As a result, a type I error occurs when the p-value is less than 0.05.

Given that we've conducted about 26 hypothesis tests, the likelihood of at least one false positive is 1 - (1 - 0.05)^26. A significance level of 0.05 means there is a 5% chance of incorrectly rejecting the null hypothesis. As a result, the more hypothesis tests, the probability of making a type I error increases.

Q11 | What does it mean to reject the null hypothesis in a chi-sqaured test of independence?

The chi-squared test of independence is a hypothesis test that compares two categorical values and measures if there is a dependence between them. The null and alternative hypothesis is as follows:

- **Null Hypothesis:** there is no association between two categorical variables
- **Alternative Hypothesis:** there is association between two categorical variables

To reject the null hypothesis in a chi-squared test of independence means to accept the alternative hypothesis, which means there is association between two categorical variables.

Q12 | Reflect on the limitations of hypothesis testing methods in general. In what ways do these methods provide insight into data, and what are their potential drawbacks or pitfalls?

Some limitations of hypothesis testing is being too dependent on the p-value, as when the p-value is less than 0.05, it simply just means that the result is statistically significant, that's all. Other drawbacks are that such hypothesis tests assume a normal distribution, and if the data isn't normally distributed, it affects the validity of the results. Lastly, causation is not correlation.

This means that hypothesis testing doesn't provide evidence for relationships between variables. Even if two variables are related, it doesn't mean that one causes the other. Hypothesis testing also involves risk of errors, such as Type I and Type II errors.

- A Type I Error occurs when the null hypothesis is wrongly rejected (false positive)
- A Type II Error occurs when the null hypothesis is wrongly accepted (false negative)

Despite the limitations, hypothesis testing methods provides insights by helping to make decisions with regards to accept or reject a hypothesis. Moreover, it allows us to quantify the strength of the evidence against the null hypothesis.

Q13 | You have dealt with two versions of the response variable – life expectancy in years (a numerical variable) and an indicator showing which counties have above average life expectancy (a binary categorical variable). Did the results of your hypothesis tests depend on which of these you used? That is, were there any predictors that had statistically significant relationships with one of these two versions of the response but not the other? How would this affect future machine learning?

When we look at the results of the hypothesis tests, we can see that despite the difference in terms of a numerical and binary categorical variable, it didn't depend on the type of the data because the results were all True with regards to being Statistically Significant. This might mean that the relationship between the predictors and life expectancy is strong enough that it is evident. A reason this could happen is because that factors such as smoking and obesity have a impact on the life expectancy regardless of whether it is categorical or numerical variable. With regards to machine learning, it depends on the following:

- If we treat life expectancy like a continuous variable, we would likely use regression models
- If we treat life expectancy like a binary variable, we would likely use classification models.

Q14 | Reflect on any challenges or insights gained during the process of conducting hypothesis tests and interpreting results. What lessons have you learned that could inform future research or data analysis endeavors

With regards to challenges and insights gained, I learned more about how hypothesis testing works, and the different ways it can be affected when using a numerical or categorical variable. I learnt a lot about the applications of it. The most valuable lesson I've learned is the importance of data preparation and understanding the assumptions of the hypothesis tests I use. Ensuring that the data is clean, that missing values are handled appropriately, and that the assumptions of the tests are checked will improve the reliability of the results. Lastly, I understand the importance of clearly defining the research question and choosing the appropriate test based on the type of data and the relationship we're investigating. In future research, I will continue to use these lessons to make informed choices and refine my approach to hypothesis testing, ensuring that the conclusions I draw are valid and meaningful.