Madeleine Sharp

DSC530 – 12.2 Assignment: Term Project

Project Topic: Colicky Horses

Introduction:

For this project, I elected to utilize a horse colic dataset to find some insights regarding the biomarkers, colic surgery status, and life outcomes of the horses present within the dataset. For some background, colic indicates a painful problem within a horse's abdomen. It is often unpredictable, frequently unpreventable, a common concern for horse owners, and horses are naturally prone to colic. However, treatment and surgery can help.

Statistical/Hypothetical Question:

For this project, I thought of two research questions I would like to explore, and one hypothesis (with a sub-hypothesis).

Research Questions:

1. What do the overall biomarkers look like for colicky horses (respiratory rate, pulse, etc.)?
2. What is the nature of the relationships between these biomarkers? Do any of the biomarkers impact a horse's life outcome more-so than others?

Hypothesis:

1. I hypothesize that horses who received surgery for colic were less likely to die or be euthanized than those that did not receive colic surgery.
    - I hypothesize that the relationship between surgery status and a horse's life outcome is a significant one.

Outcome of EDA:

The outcome of the EDA indicated to me that, overall, my dataset did not include really any outliers. This could be because there were a predominant number of missing values, which I subsequently handled depending upon variable type. Given this, any missing values were filled with either a mean (numeric variables) or a mode (categorical variables), and both of those values would fall within the range of values present, not outside of those (since they are measures of central tendency).

Nearly all of my true numeric variables followed a specific type of distribution (either normal or right-skewed). The categorical variables did not follow a specific distribution, but this is understandable – these variables did not have a true range of integer values, but rather numeric representations of fixed categories.

From the remainder of my EDA, it did not appear that the biomarker variables had strong relationships with one another, at least not as much as I thought they might. For my PMFs, I gleaned that perhaps these visualizations indicated that horses with higher pulses were deemed to be better candidates to receive surgery, given their pulses were higher, and thus a higher pulse could indicate systemic distress that warranted intervention. For my CDF, I was able to determine that overall, most of the horses did not have severe abdominal distension, in fact, the combined groups of horses that had no, slight, and moderate distension made up approximately 85% of the dataset. From my scatterplots, it appeared that a slight linear relationship existed between pulse and respiratory_rate, although the values were largely concentrated in one area before dispersing a bit. The strongest, positive correlation existed between packed_cell_volume and mucous_membrane. I was also able to find that while the correlation between surgery status and life outcome was a positive, weak relationship, the p-value indicated statistical significance.

Additional Thoughts:

From this analysis, I feel as though I could have done some additional analyses by utilizing some of the other variables present within the dataset that I dropped. I also believe that having more datapoints (instead of just 299) may have been helpful and bolstered the sample size. Also, given about 30% of the data was missing from the original dataset, I had to handle those missing values, and that could have impacted the analysis outcomes. Additionally, some of the variables, such as capillary_refill_time, would have been better if they were not categorical and had exact integer measures.

I assumed that the relationship between a horse's surgery option and its life outcome would be stronger than it was. Overall, I did not seem to really find much in the way of relationships that garnered any substance/merit. From a challenges standpoint, I changed my project topic twice before landing on this, as I had some issues finding a good dataset that would work for the objectives and scope of this project. My main takeaway from this project is that I would like to move forward and obtain additional practice utilizing these measures from this course and that we learned in our ThinkStats text – both within my education and out in the real world. I really enjoyed using a dataset that focused on a topic I was passionate about for this project to learn more and see if any significant insights existed.