For this project, I used a data set of colic in horses that I found at the following link[[1]](#footnote-1). There are two data sources that I used from this link, the csv file containing the data set and a txt file that is a data dictionary containing fuller descriptions of the data. Colic in horses can vary greatly in severity.[[2]](#footnote-2)

Below are the research questions that I focused on:

* Is there a correlation between any of the variables?
* Are there any outliers?
* Are there any major differences observed in horses that survived and horses that did not?
* Is there a statistically significant relationship between any of the variables and the outcome?

I looked at the correlation between total protein and packed cell volume and it is quite a weak correlation. When looking at the correlation between pulse and respiratory rate, there was a moderate, positive correlation. From looking at the histograms of each variable, rectal temperature has the most normal bell-shaped gaussian distribution while all others appeared slightly right skewed. Respiratory rate and packed cell volume were bimodal. Each variable was sorted to view the ten highest and smallest values in order to identify any outliers and none were identified or removed. From looking at horses that survived and horses that did not survive, those that did not survive were observed to have a higher probability of having a higher packed cell volume. 20% of rectal temperatures were below 38.0 °C and 90% of rectal temperatures were below 39.0 °C, which is interesting because our mean for this variable is 38.2. While pulse appears to have a CDF with an exponential pattern, when plotting the complementary CDF, it actually does not.

The null hypothesis used for the hypothesis testing is that there is no difference in packed cell volume between horses that survived and horses that did not survive. When testing the difference in mean, there was no intersection observed at the CDF and the p-value was 0.0, meaning the difference in packed cell volume is statistically significant.

For multiple regression, the dependent variable used was packed cell volume and the explanatory variables were the outcome of the horse surviving, respiratory rate and pulse. All p-values, except respiratory rate, were below 0.05 indicating a statistically significant relationship.

Overall, I think it would have been helpful to have more context about the variables, such as how the data was obtained. Were the horses medically sedated prior? As most of the data is medical, was the testing done at specific intervals? For this project, I struggled with determining which variables to use for what requirement. I wanted to apply everything (CDFs, PMFs, regression) to all of the variables, but did my best to follow the assignment instructions.

1. Uci. (2017, June 06). Horse Colic Dataset. Retrieved May 28, 2020, from <https://www.kaggle.com/uciml/horse-colic> [↑](#footnote-ref-1)
2. PetMD. (2019, December 20). Colic in Horses: Signs, Causes and Treatment. Retrieved from <https://www.petmd.com/horse/conditions/digestive/c_hr_equine_colic> [↑](#footnote-ref-2)