COLICKY HORSES

DSC530 Winter ‘21/’22 Term Project
Madeleine Sharp

BACKGROUND: COLIC

- Colic indicates a painful problem within a horse's abdomen. Because colic is often unpredictable and frequently unpreventable, it's a common concern for horse owners. Horses are naturally prone to colic, however, treatment and surgery can help (horses are physically unable to vomit, so any abdominal/digestive issue a horse experiences, such as colic, is great cause for concern and intervention).
- The most common types of colic are related to impaction, in which undigested feed or foreign bodies such as parasites block the movement of digesta through the intestines and cecum.

RESEARCH QUESTIONS

- Two research questions I would like to consider for my project include:

  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

- One hypothesis that I would like to consider/evaluate is:
    1. I hypothesize that horses of which 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.

DATA SOURCE

- The dataset I am using for the project is a flat-file source (.csv file) and the data can be obtained/accessed from the below places:
  1. https://archive.ics.uci.edu/ml/datasets/Horse+Colic
  2. https://www.kaggle.com/uciml/horse-colic?select=datadict.txt

DATASET VARIABLES

- The original dataset includes a total of 28 variables.
- I will be using 11 of those variables for my project.
- See the variables below (descriptions on next slide):
  - surgery
  - age
  - rectal_temp
  - pulse
  - respiratory_rate
  - mucous_membrane
  - capillary_refill_time
  - abdominal_distension
  - abdomen
  - packed_cell_volume
  - outcome

DATASET VARIABLES: DESCRIPTIONS

- **surgery:** a variable that indicates whether a horse had colic surgery or not (categorical).
  - 1 = no Surgery
  - 2 = surgery
- **age:** the age of the horse, split into two (2) categories.
  - 1 = young horse (< 6 months)
  - 2 = adult horse
- **rectal_temp:** the rectal temperature of the horse (in degrees Celsius). Normal is 37.8.
- **pulse:** the horse's heart rate in beats per minute. Normal range (for adults) is 30-40 bpm.
- **respiratory_rate:** breaths per minute of a horse. Normal range is 8-10 breaths.
- **mucous_membrane:** color of these membranes indicates current circulation. Normal pink and bright pink indicate normal or slightly increased circulation.
- **capillary_refill_time:** another indicator of circulation. The longer the refill, the poorer the circulation.
  - 1 = less than 3 seconds
  - 2 = more than or equal to 3 seconds

DATASET VARIABLES: DESCRIPTIONS CONTINUED

- **abdominal_distension:** an important parameter/indicator for colic (since colic is an abdominal/digestive issue). No distension is healthy/normal.
  - 1 = none
  - 2 = slight
  - 3 = moderate
  - 4 = severe
- **abdomen:** the status of the horse's abdomen overall. A value of 3 is likely an obstruction caused by mechanical impaction. 4 and 5 indicate a surgical lesion.
  - 1 = normal
  - 2 = other
  - 3 = firm feces in the large intestine
  - 4 = distended small intestine
  - 5 = distended large intestine
- **packed_cell_volume:** the number of red cells by volume in the blood, normal range is 30-50.
- **outcome:** the life outcome of the horse – did it live, die, or was it euthanized?

COLICKY HORSES: BACKGROUND

Overall, some additional background regarding colicky horses and the variables in this dataset includes:

- A horse's rectal temperature:
  - An elevated temp may occur due to infection.
  - Temperature may be reduced when the animal is in late shock
  - This parameter will usually change as the problem progresses
- A horse's pulse:
  - It is rare to have a lower-than-normal rate although athletic horses may have a rate of 20-25
  - Those with painful lesions or suffering from circulatory shock may have an elevated heart rate
- A horse's mucous membrane:
  - 1 and 2 probably indicate a normal or slightly increased circulation (normal pink, bright pink)
  - 3 may occur in early shock (pale pink)
  - 4 and 6 are indicative of serious circulatory compromise (pale cyanotic, dark cyanotic)
  - 5 is more indicative of a septicemia (bright red/injected)
- A horse's abdominal distension:
  - Abdominal distension is likely to be painful and have reduced gut motility
  - A horse with severe abdominal distension is likely to require surgery just to relieve the pressure
- A horse's packed cell volume:
  - The level rises as the circulation becomes compromised or as the animal becomes dehydrated

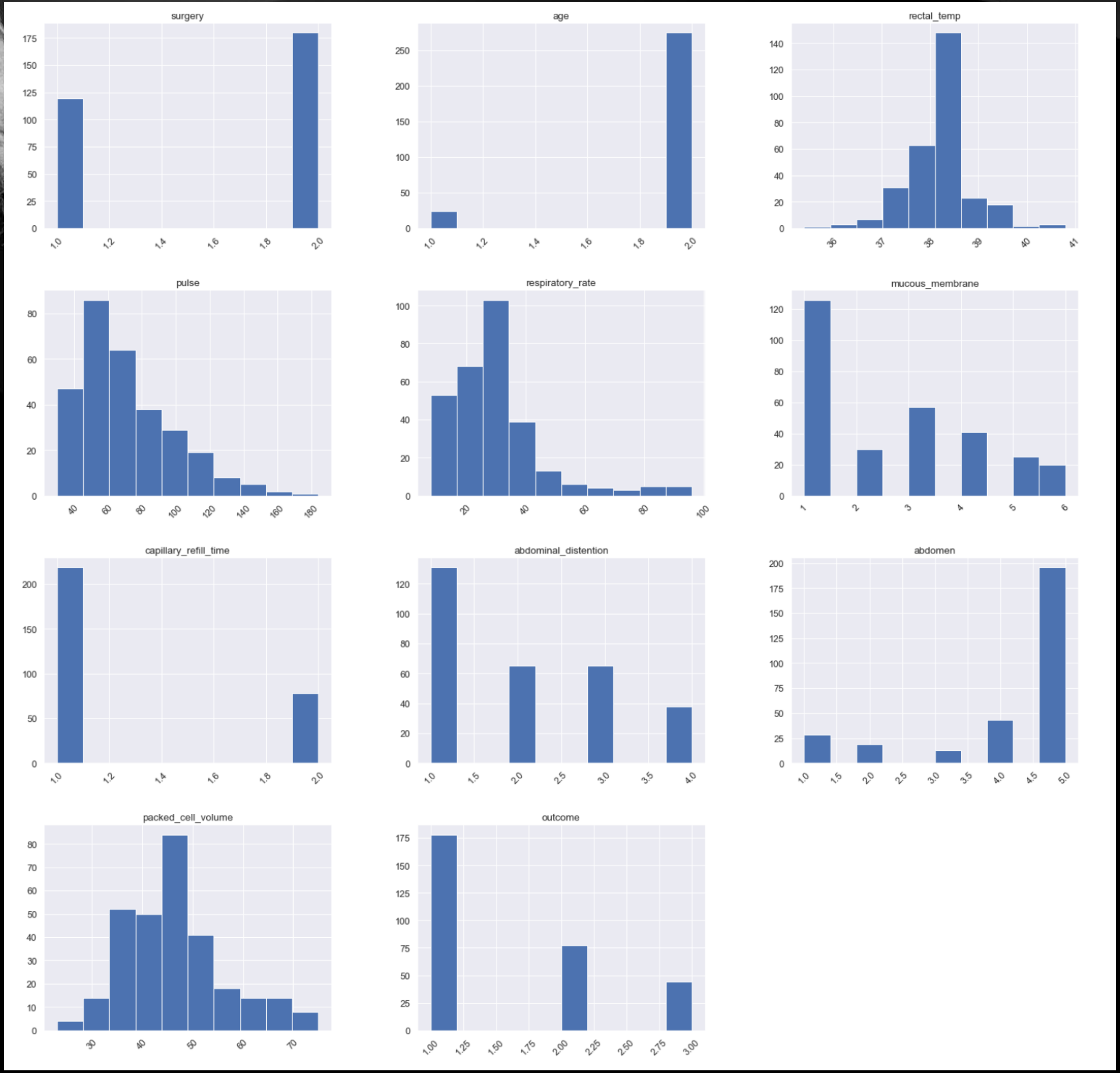
DATA CLEANING

- Cleaning my dataset was imperative so that I could use the data in the ways I need for this project.
- These steps included:
  - Import the data
  - Drop any variables I wasn’t using
  - Find and replace missing values
    - 30% of the values were missing from this dataset, which is significant.
    - Rather than remove all values and severely cut down the data, I elected to fill missing values with a measure of central tendency:
      - Mean for numeric variables
      - Mode for categorical variables
  - Transform categorical variables so they could be utilized in analyses and measures.
    - Checking data types.
    - Assigning a numeric encoded value to each categorical variable.
      - Then converting to a float to be used in analyses.

PROJECT IMPACT OF VARIABLES

- Each of these variables will impact my project research questions because:
  - These variables are biomarkers that give information about a horse's health, specifically those horses that have colic.
  - Because colic is considered an illness, and when a living being is ill, biomarkers can be impacted by that and can be telling as to something being wrong with the living being (in this case, a horse).
  - In addition to biomarker variables, I also have variables that indicate whether a horse has received colic surgery and their age - these are variables which may impact a horse's life outcome.
  - Lastly, knowing the outcome variable alongside the other variables may grant insight into the potential relationships between surgery, age, biomarkers, and life outcome, etc.

HISTOGRAMS


surgery
175 150 125 100 75 50 25 0
1.0 1.2 1.4 1.6 1.8 2.0
age
250 200 150 100 50 0
1.0 1.2 1.4 1.6 1.8 2.0
rectal_temp
140 120 100 80 60 40 20 0
36 37 38 39 40 41
pulse
80 60 40 20 0
40 60 80 100 120 140 160 180
respiratory_rate
100 80 60 40 20 0
20 40 60 80 100
mucous_membrane
120 100 80 60 40 20 0
1 2 3 4 5 6
capillary_refill_time
200 150 100 50 0
1.0 1.2 1.4 1.6 1.8 2.0
abdominal_distention
120 100 80 60 40 20 0
1.0 1.5 2.0 2.5 3.0 3.5 4.0
abdomen
200 175 150 125 100 75 50 25 0
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
packed_cell_volume
80 70 60 50 40 30 20 10 0
30 40 50 60 70
outcome
175 150 125 100 75 50 25 0
1.00 1.25 1.50 1.75 2.00 2.25 2.50 2.75 3.00

HISTOGRAM SUMMARY + ANALYSIS

- surgery:
  - Categorical variable distribution.
  - We can see that from the dataset, more horses did have surgery, and those that did not have surgery were about 3/4 or so of the amount of those that did.
- age:
  - Categorical variable distribution.
  - We can see that from the dataset, drastically more adult horses are present within this data, with only about 25 young horses.
- rectal_temp:
  - Overall, we can see that rectal temperature follows a pretty normal distribution curve.
- pulse:
  - This distribution takes the shape of right-skewed (tails off to the right) distribution - so more of the pulse values are heavily concentrated at the beginning of the distribution.
- respiratory_rate:
  - This distribution is also quite right-skewed.

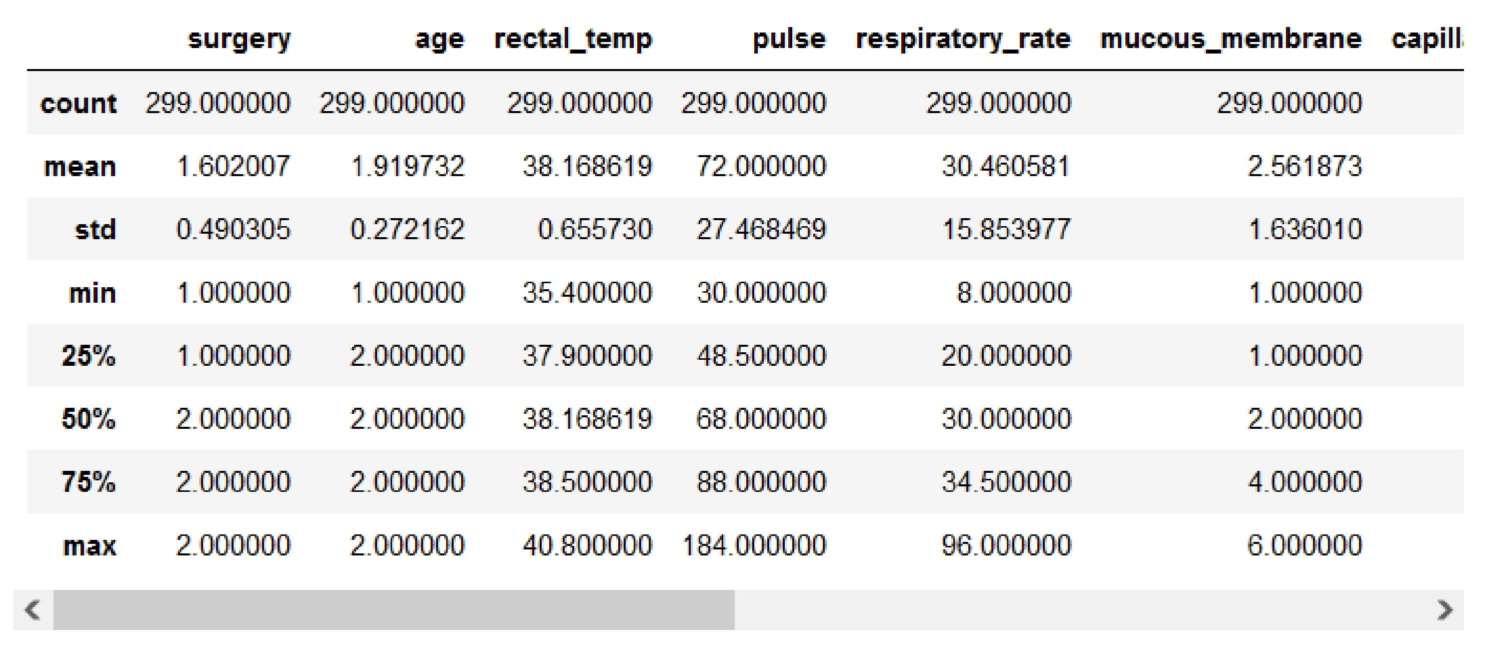
HISTOGRAM SUMMARY + ANALYSIS CONTINUED

- mucous_membrane:
  - We can see that from the dataset, most horses' mucous membranes were normal pink in color (considered healthy).
- capillary_refill_time:
  - We can see more horses had a capillary refill time of less than 3 seconds.
- abdominal_distension:
  - Most horses' abdominal distension was that they had none - a healthy sign.
- abdomen:
  - We can see that from the dataset, most horses had a distended large intestine.
- packed_cell_volume:
  - Overall, we can see that packed cell volume follows a pretty normal distribution curve.
- outcome:
  - We can see that from the dataset, most horses ended up living.

HISTOGRAM SUMMARY + ANALYSIS OUTLIERS

- Overall, my dataset did not include many of outliers. This may be because there were a predominant number of missing values, of which I have handled appropriately depending on the variable type. Overall, the data points for each of the variables chosen really did fall within range, and therefore are usable within that range - no need to remove anything since no prevalent outliers were truly present.

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES



	surgery	age	rectal_temp	pulse	respiratory_rate	mucous_membrane	capill
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000	
mean	1.602007	1.919732	38.168619	72.000000	30.460581	2.561873	
std	0.490305	0.272162	0.655730	27.468469	15.853977	1.636010	
min	1.000000	1.000000	35.400000	30.000000	8.000000	1.000000	
25%	1.000000	2.000000	37.900000	48.500000	20.000000	1.000000	
50%	2.000000	2.000000	38.168619	68.000000	30.000000	2.000000	
75%	2.000000	2.000000	38.500000	88.000000	34.500000	4.000000	
max	2.000000	2.000000	40.800000	184.000000	96.000000	6.000000	

This code output table is available within my code output in its entirety.
The above is merely a snapshot.

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON’T.

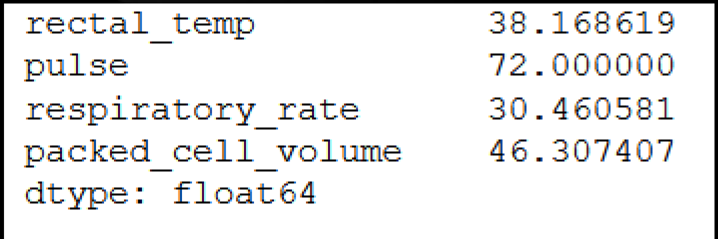
- The table on the previous slide shows a summary of some basic descriptives of this dataset. In this, we can see:
  - count: this is the number of instances (data points) in the dataset for each variable. Because missing values were handled, each variable has the same count number.
  - mean: this is the average of the values for each of the variables. However, this can be somewhat misleading with certain variables, which is why further analyses will take place below in subsequent code steps.
  - std: this is the standard deviation of each variable.
  - min: this is the smallest value for each variable within the dataset.
  - interquartile ranges (25%, 50%, 75%): this is a measure of statistical dispersion, or spread.
  - max: this is the largest value for each variable within the dataset.

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON'T.

- While the previous table and descriptions are an initial overview of descriptive characteristics of this dataset, I will delve further into these variables.
- This is because, while the output table shows means for each variable, for example, this is not actually helpful in the cases of those categorical variables that have simply been assigned numeric values for each category. For true numeric variables, I will keep the means, but for categorical variables, we will find the mode.

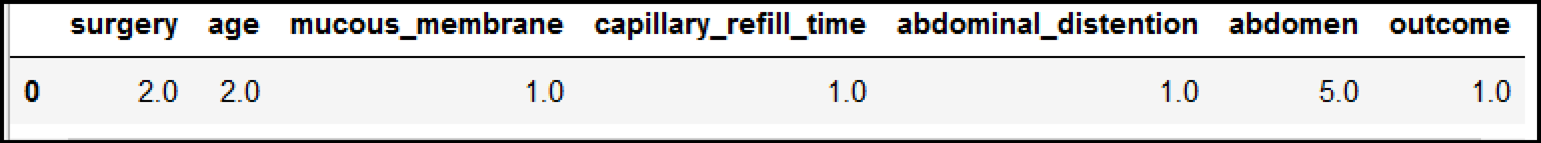
DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON'T.

- Means (for true numeric variables):



```
rectal_temp           38.168619
pulse                 72.000000
respiratory_rate      30.460581
packed_cell_volume    46.307407
dtype: float64
```

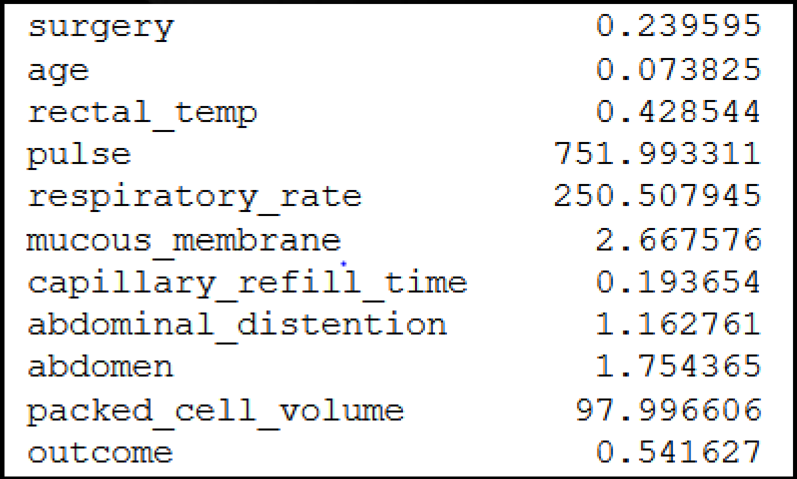
- Modes (for true categorical variables):



	surgery	age	mucous_membrane	capillary_refill_time	abdominal_distention	abdomen	outcome
0	2.0	2.0	1.0	1.0	1.0	5.0	1.0

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON'T.

- Variance (for all variables):

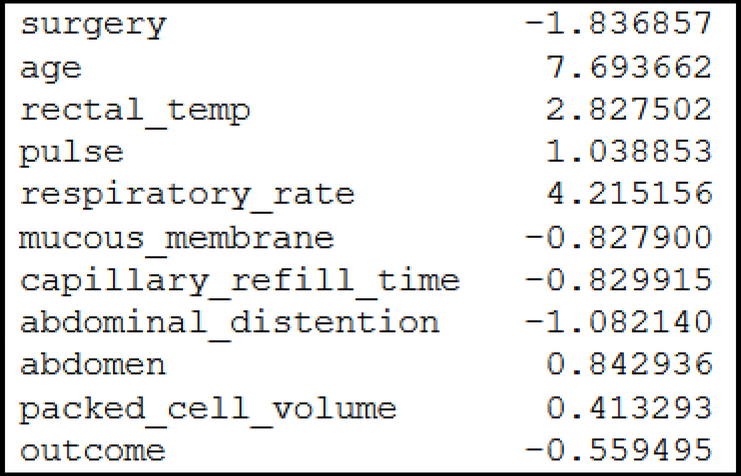


```
surgery                      0.239595
age                          0.073825
rectal_temp                  0.428544
pulse                      751.993311
respiratory_rate           250.507945
mucous_membrane              2.667576
capillary_refill_time        0.193654
abdominal_distention         1.162761
abdomen                      1.754365
packed_cell_volume          97.996606
outcome                      0.541627
```

- The variance is a measure that is used to quantify the amount of variation of a set of data values from its mean. The variance values for each of the variables is outlined in the output above.

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON'T.

- Kurtosis (for all variables):

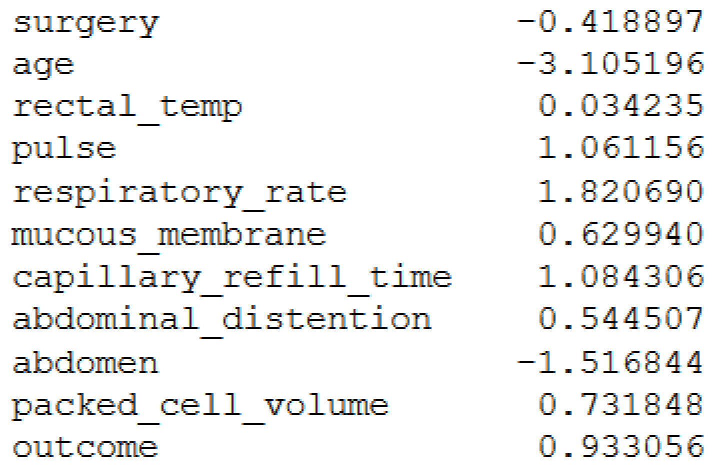


```
surgery                  -1.836857
age                       7.693662
rectal_temp               2.827502
pulse                     1.038853
respiratory_rate          4.215156
mucous_membrane          -0.827900
capillary_refill_time    -0.829915
abdominal_distention     -1.082140
abdomen                   0.842936
packed_cell_volume        0.413293
outcome                  -0.559495
```

- Kurtosis refers to one of the two measures that quantify shape of of a distribution and it describes the peakedness of the distribution. The kurtosis values for each of the variables are present above. To interpret kurtosis, the following can be adhered to:
  - For kurtosis, the general guideline is that if the number is greater than +1, the distribution is too peaked.
  - Likewise, a kurtosis of less than –1 indicates a distribution that is too flat.
- Obviously, for some of the categorical variables, kurtosis will not apply as well as for those variables that are truly numeric in nature.

DESCRIPTIVE CHARACTERISTICS OF DATA VARIABLES CON'T.

- Skew (for all variables):



```
surgery                   -0.418897
age                       -3.105196
rectal_temp                0.034235
pulse                      1.061156
respiratory_rate           1.820690
mucous_membrane            0.629940
capillary_refill_time      1.084306
abdominal_distention       0.544507
abdomen                   -1.516844
packed_cell_volume         0.731848
outcome                    0.933056
```

- Skew refers to a measure of assymmetry or distortion of symmetric distribution.
- For interpreting skewness, the rules of thumb are:
  - If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.
  - If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.
  - If the skewness is less than -1 or greater than 1, the data are highly skewed.
  - If the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed.
  - If the peak is toward the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed.

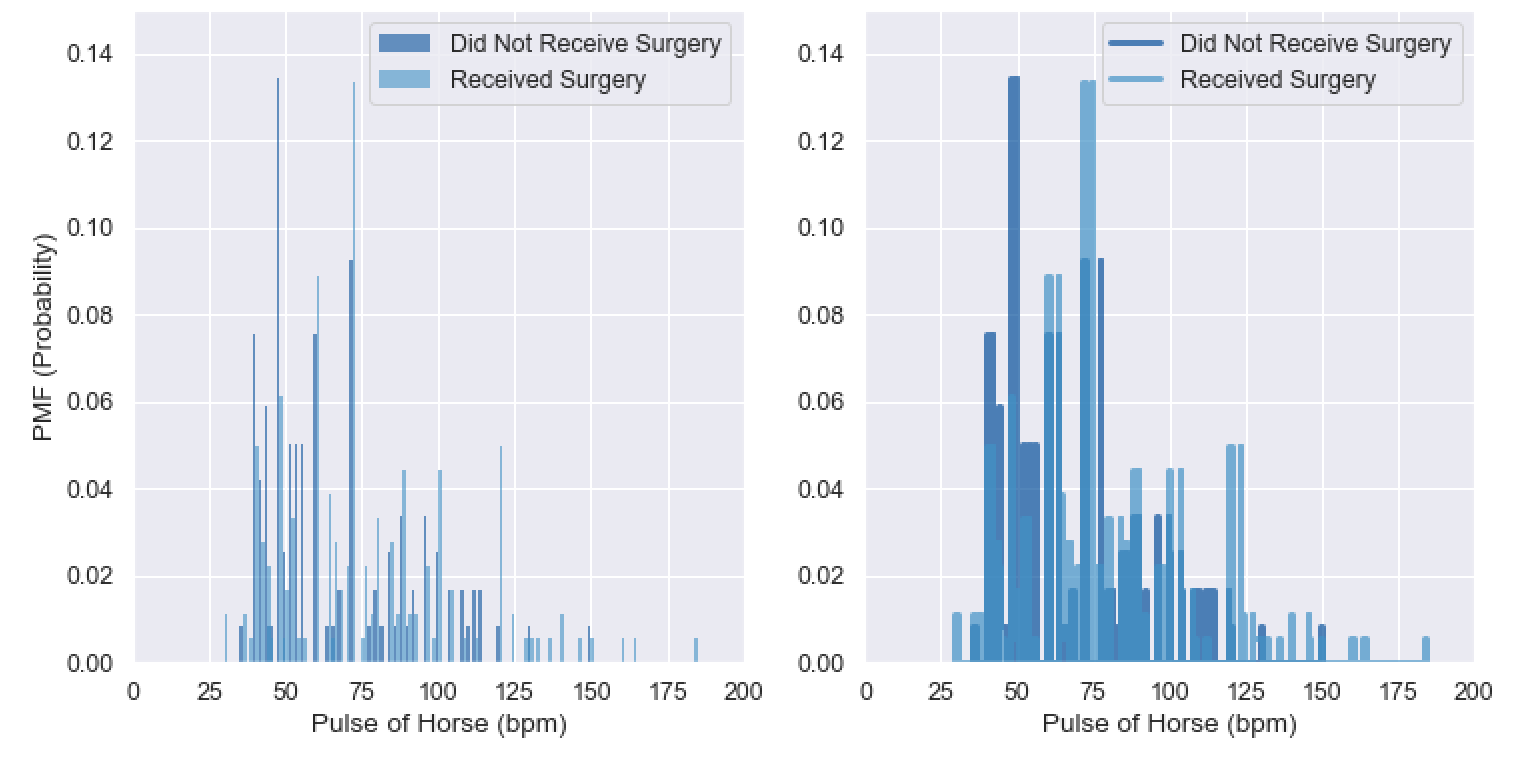
PMF

- For my PMF, I am going to explore the surgery variable alongside the pulse variable. More specifically, the variable that I am going to "slice" like a filter will be the surgery variable. I am going to compare those horses that had colic surgery to those that did not have colic surgery.

- Within this, I am going to compare, in particular, the pulses of those horses that had surgery versus the pulses of those horses that did not undergo surgery.

- I would like to explore/assess the pulse distributions of each of these groups.

PMF CONTINUED



PMF CONTINUED

- PMFs are useful because they allow us to compare the two distributions of this data without being misled by the difference in sample size. These measures map from each value to its probability.

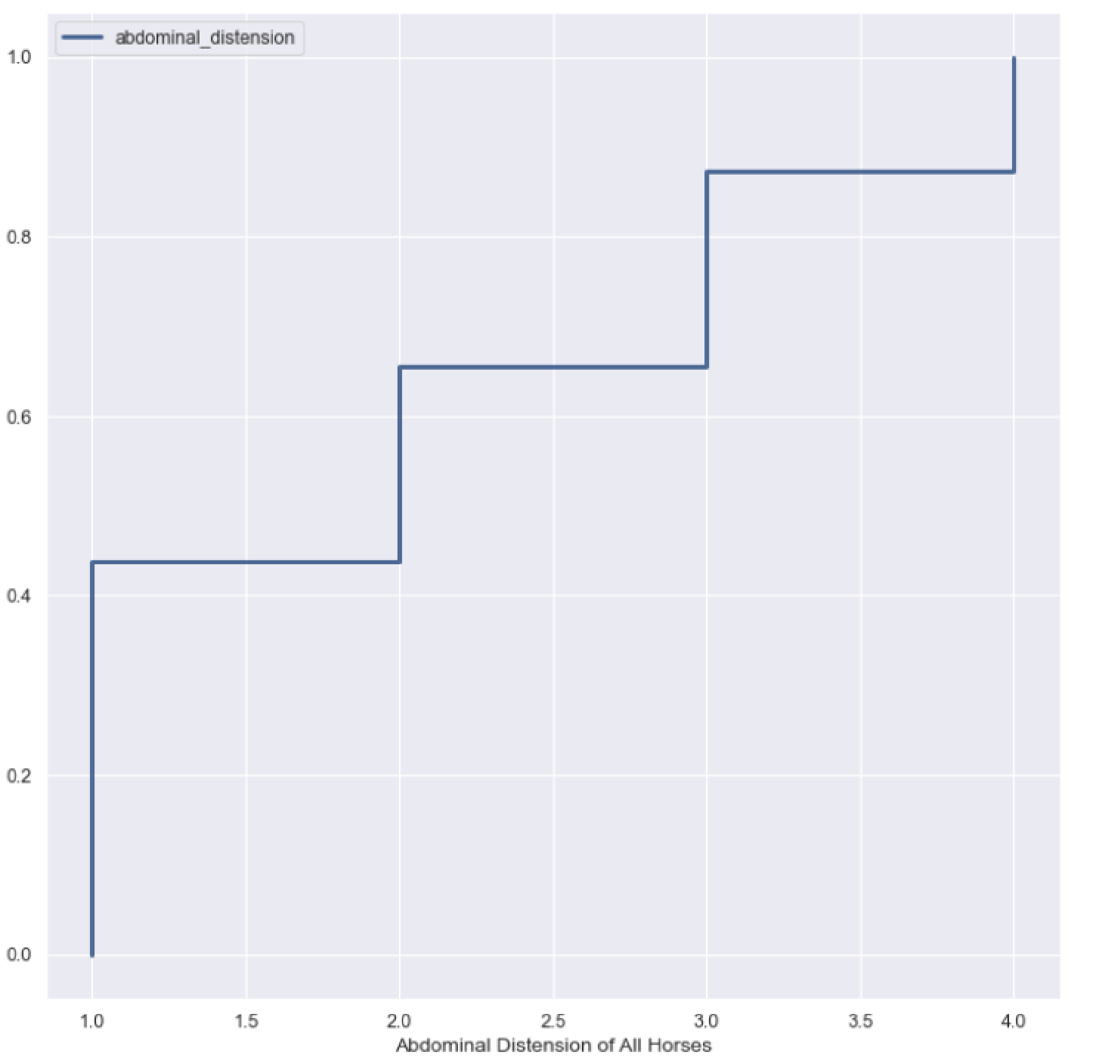
- The left figure shows the PMFs as a histogram, the right shows a step feature.
- Based on these figures, we can ascertain that those horses that did not receive colic surgery tend to have lower pulses overall (per the dark blue concentrations between about 40 bpm and 100 bpm) than those that did receive surgery (the light blue concentrated more between about 60 bpm and 140 bpm).

- This may perhaps indicate to us that the 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.

CDF

- For creating a CDF, I am going to use the variable abdominal distension. I would like to explore the percentile ranks of the status of a horse's abdominal distension. This variable can tend to be quite important when assessing a colicky horse.

CDF CONTINUED



CDF CONTINUED

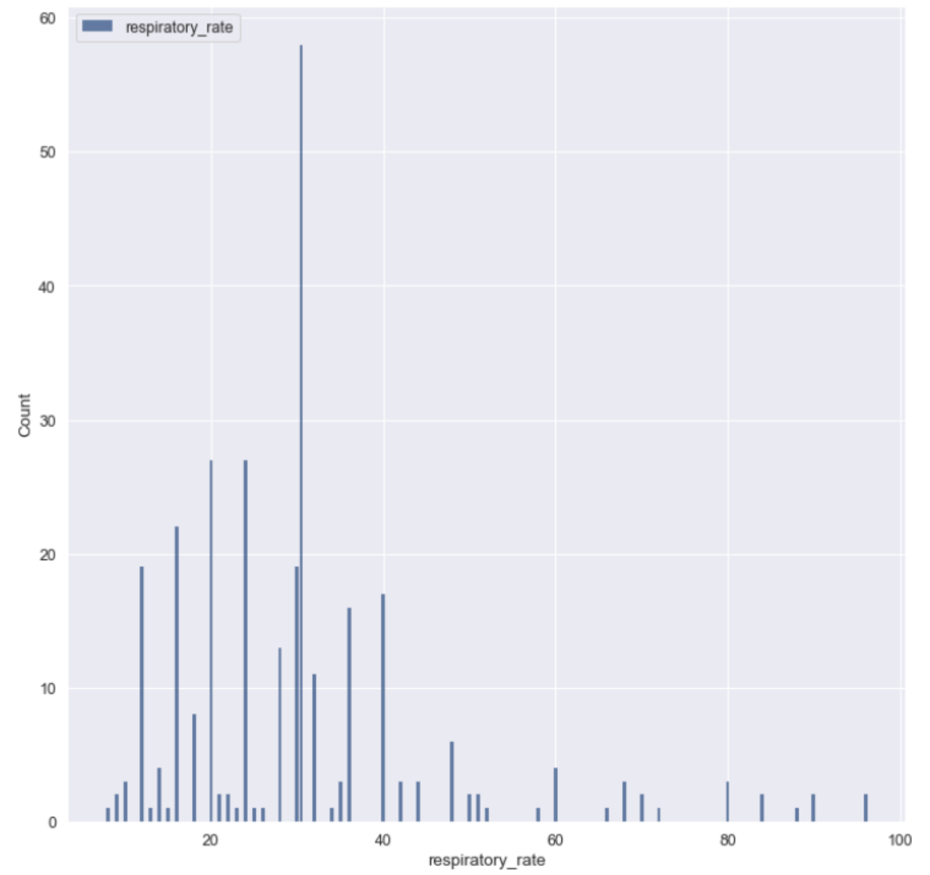
- CDFs are useful because they are a function that maps from a value to its percentile rank, which is helpful for knowing which percentile a certain value falls within.
- In my specific example above with abdominal distension, we can see that:
  - The abdominal distension group for 1 (none) is in the 45th percentile, approximately.
  - The group for 2 (slight) is in the 65th percentile, approximately.
  - The group for 3 (moderate) is in the 85th percentile, approximately/
  - Finally, the group for 4 (severe) is in the 100th percentile.
- Overall, this means that most of the horses did not have severe abdominal distension, in fact, the combined groups of horses that had no, slight, and moderate distension makes up approximately 85% of the dataset.

ANALYTICAL DISTRIBUTION/MODELING DISTRIBUTIONS

- For the analytical distribution, I will be plotting a Normal (Gaussian) Distribution of horse respiratory rates. This will show me the overall curve of the distribution of colicky horse's respiration. I will also complete a Normal Distribution using the CDF for this variable.

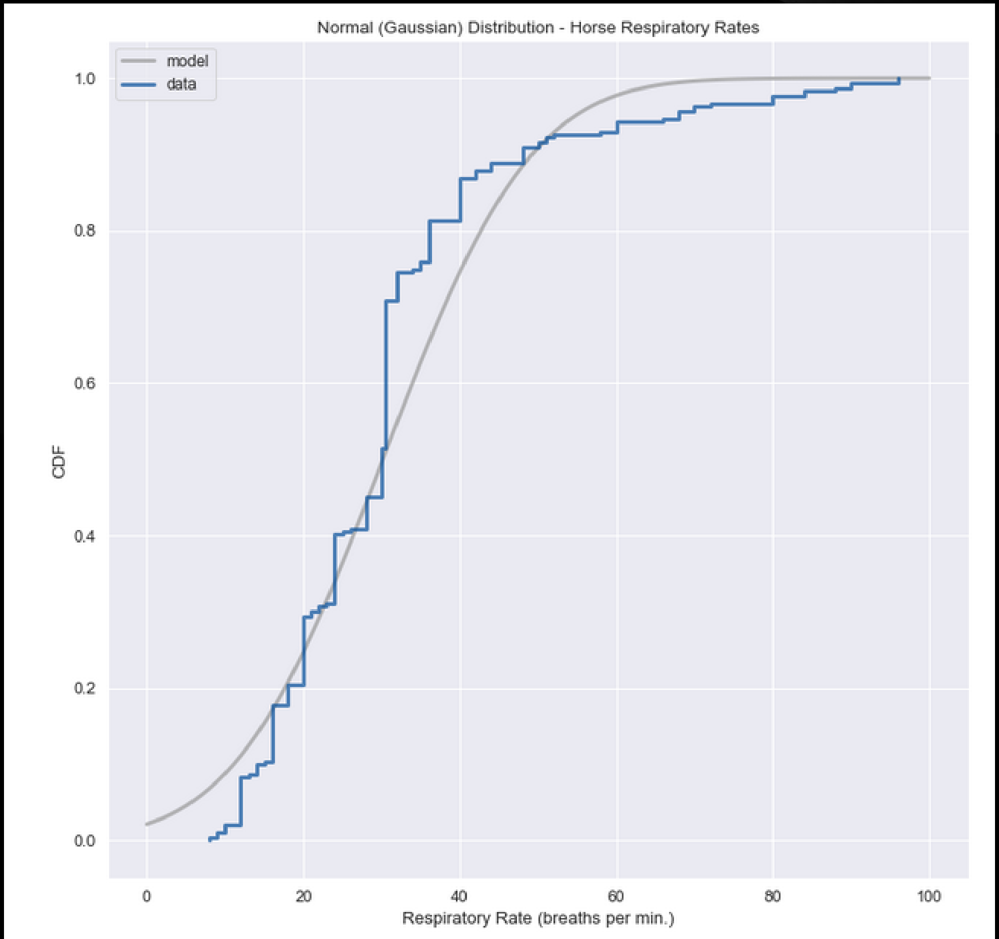
ANALYTICAL DISTRIBUTION/MODELING DISTRIBUTIONS

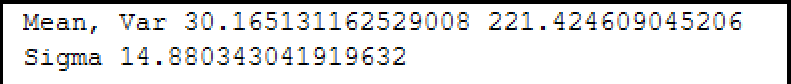
Normal Distribution (also present in my earlier histograms)



ANALYTICAL DISTRIBUTION/MODELING DISTRIBUTIONS CON'T.

Normal Distribution with CDF





```
Mean, Var 30.165131162529008 221.424609045206
Sigma 14.880343041919632
```

ANALYTICAL DISTRIBUTION/MODELING DISTRIBUTIONS CON'T.

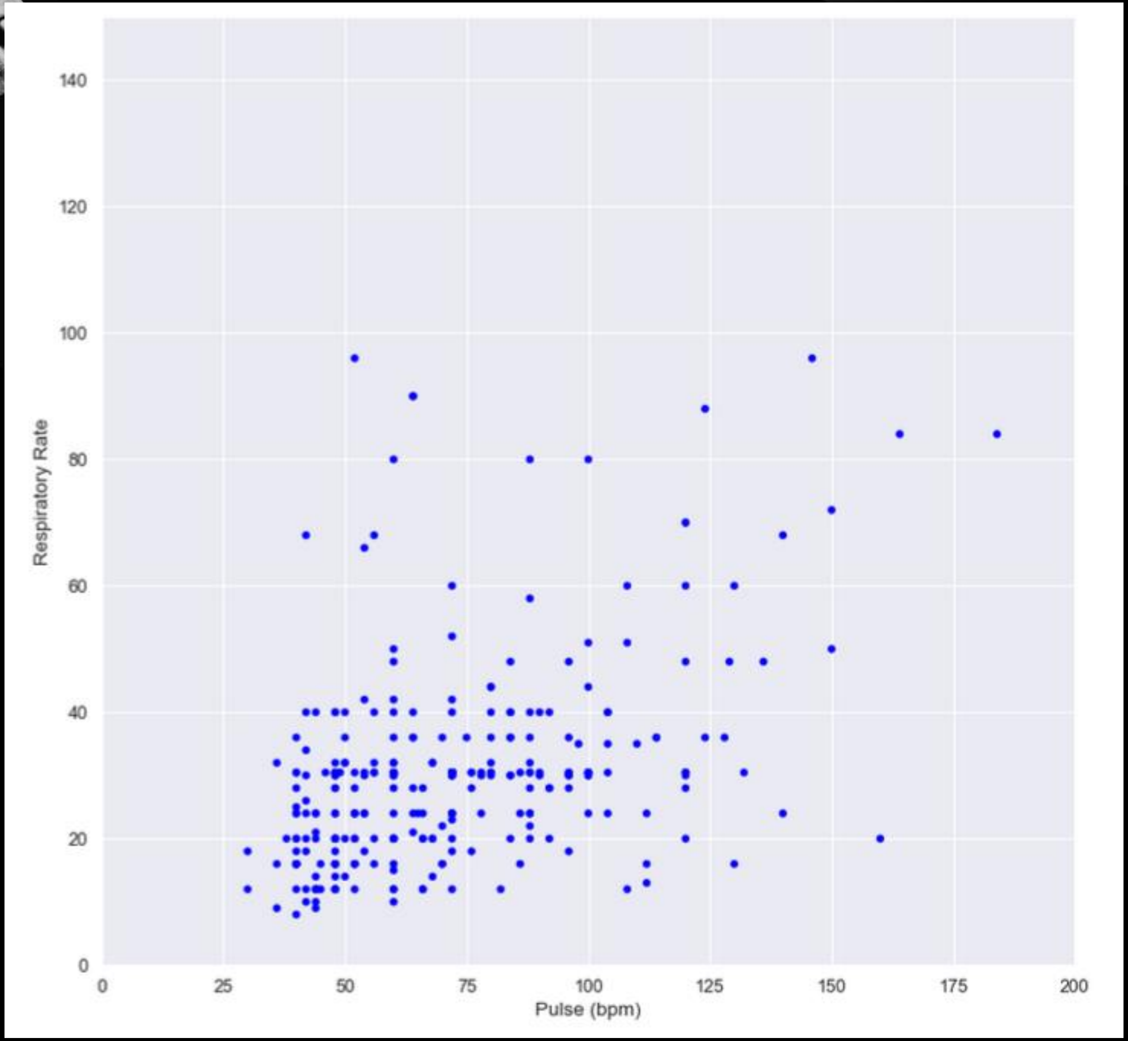
- In the main normal distribution, we can see what we were able to see in the histogram EDA towards the beginning of this file.
- In the distribution with CDF, we can see that while the model overall seems to follow the curve of the actual data, there are a few places in which there is a discrepancy, and it falls out of line. These places are:
  - Below the 20th percentile of the data.
  - Between the 55th and 90th percentiles (it briefly touches right at the 90th percentile).
  - And finally, from the 90th percentile to the 100th (also briefly touches on the 100th percentile).
- From this, we can also see that the actual data values do not produce the straightest line - lots of jaggedness in that line there. This also merely means that the normal model may not be the best fit for this data.

SCATTERPLOTS

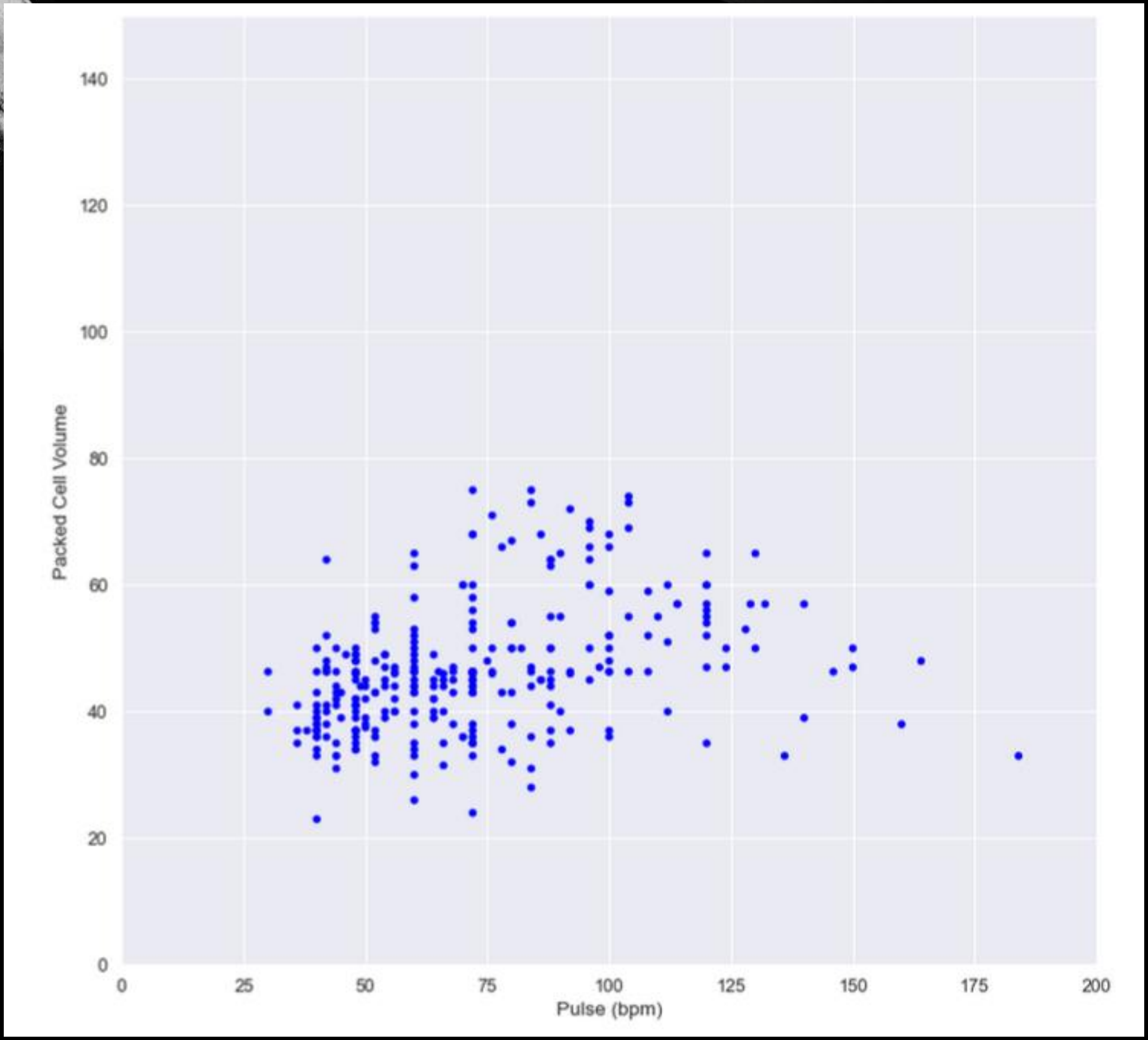
- The first two variables that I will compare in my first scatterplot are pulse and respiratory_rate.

- The next two variables that I would like to assess in my second scatterplot are pulse and packed_cell_volume.

SCATTERPLOTS CONTINUED

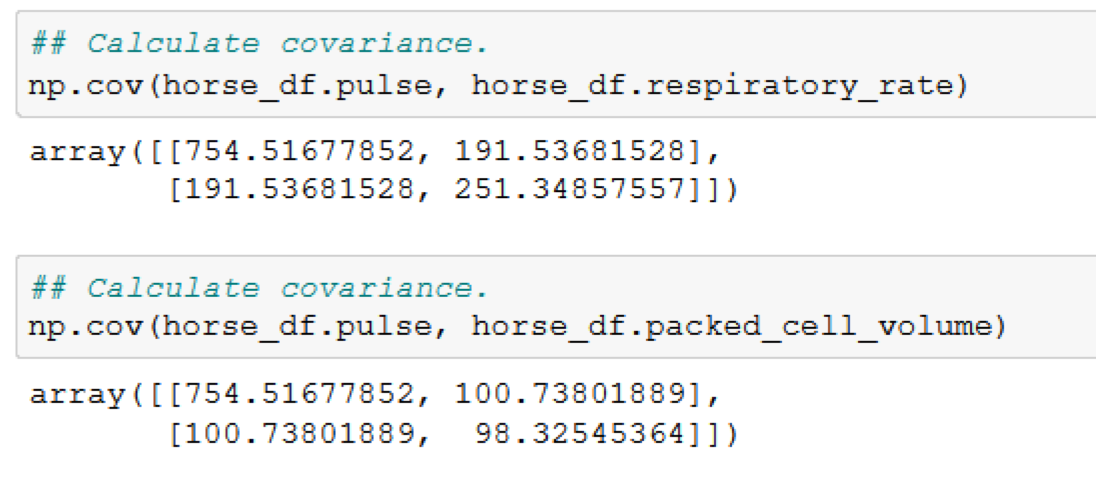


SCATTERPLOTS CONTINUED



SCATTERPLOTS CONTINUED

- The calculations for the covariance of the two scatterplot relationships are below.



```
## Calculate covariance.
np.cov(horse_df.pulse, horse_df.respiratory_rate)
```

```
array([[754.51677852, 191.53681528],
       [191.53681528, 251.34857557]])
```

```
## Calculate covariance.
np.cov(horse_df.pulse, horse_df.packed_cell_volume)
```

```
array([[754.51677852, 100.73801889],
       [100.73801889,  98.32545364]])
```

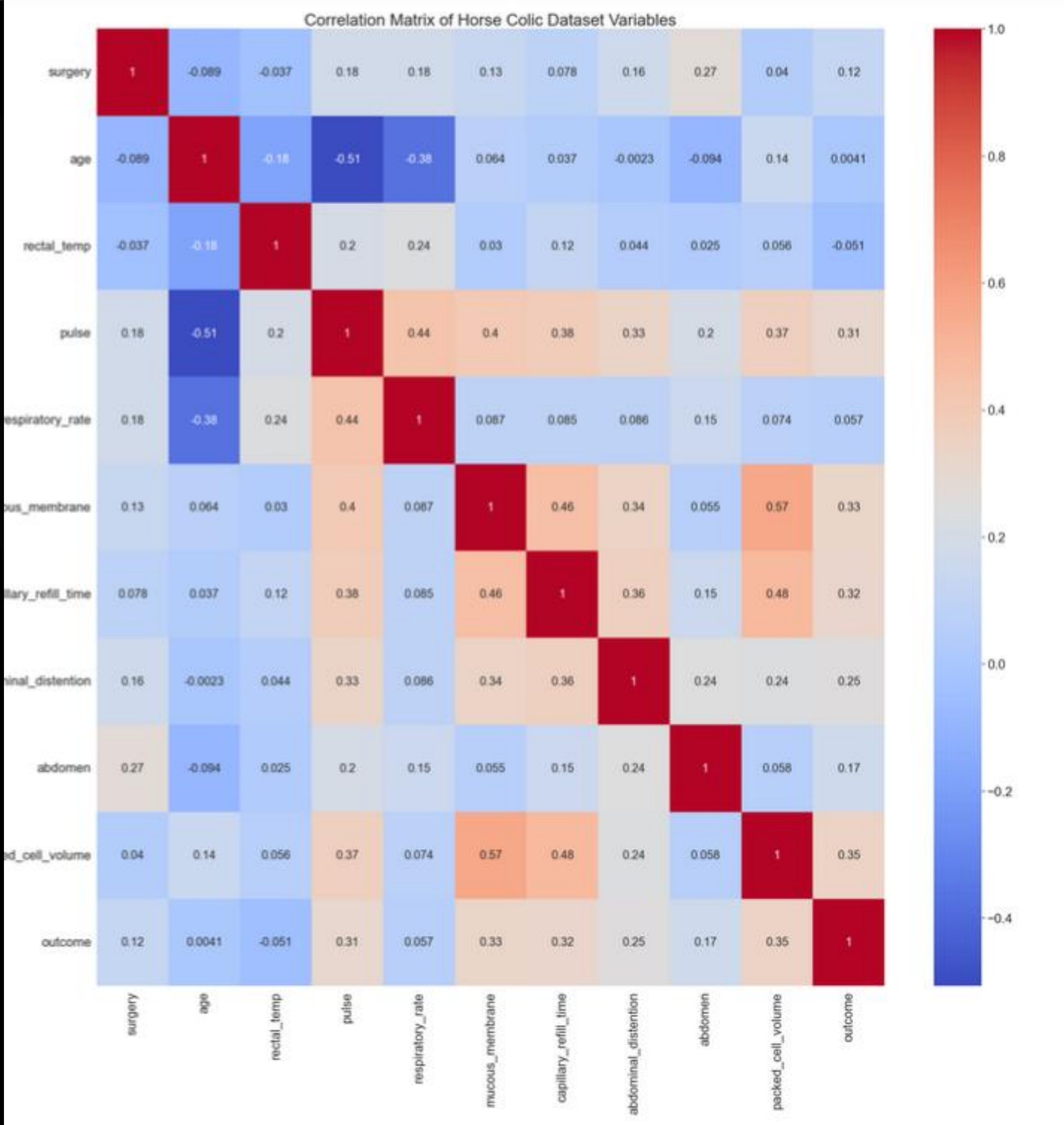
SCATTERPLOTS CONTINUED

- The nature of the relationships between the variables in the scatterplots:
  - For the first scatterplot, we can ascertain that it looks like a linear relationship - the scatterplot follows a "line" from the bottom left of the graph up towards the top right.
  - For the second scatterplot, we see that is follows no real line, and rather almost has sort of a curve that dips downward on the right side. This scatterplot has the characteristics of a more non-linear relationship.

SCATTERPLOTS CONTINUED

- In addition to the scatterplots for looking at relationships, I wanted to complete a correlation heatmap (see next slide) to get a look at the relationships between all of the variables, since that is a part of my research question. This style of heatmap portrays the Pearson's correlation coefficient value for/between all of the variables - it grants a nice overview of the relationships between the features of the entire dataset.

SCATTERPLOTS CONTINUED



Correlation Matrix of Horse Colic Dataset Variables

	surgery	age	rectal_temp	pulse	respiratory_rate	mucous_membrane	capillary_refill_time	abdominal_distention	abdomen	packed_cell_volume	outcome
surgery	1	-0.089	-0.037	0.18	0.18	0.13	0.078	0.16	0.27	0.04	0.12
age	-0.089	1	-0.18	-0.51	-0.38	0.064	0.037	-0.0023	-0.094	0.14	0.0041
rectal_temp	-0.037	-0.18	1	0.2	0.24	0.03	0.12	0.044	0.025	0.056	-0.051
pulse	0.18	-0.51	0.2	1	0.44	0.4	0.38	0.33	0.2	0.37	0.31
espiratory_rate	0.18	-0.38	0.24	0.44	1	0.087	0.085	0.086	0.15	0.074	0.057
ous_membrane	0.13	0.064	0.03	0.4	0.087	1	0.46	0.34	0.055	0.57	0.33
llary_refill_time	0.078	0.037	0.12	0.38	0.085	0.46	1	0.36	0.15	0.48	0.32
inal_distention	0.16	-0.0023	0.044	0.33	0.086	0.34	0.36	1	0.24	0.24	0.25
abdomen	0.27	-0.094	0.025	0.2	0.15	0.055	0.15	0.24	1	0.058	0.17
d_cell_volume	0.04	0.14	0.056	0.37	0.074	0.57	0.48	0.24	0.058	1	0.35
outcome	0.12	0.0041	-0.051	0.31	0.057	0.33	0.32	0.25	0.17	0.35	1

SCATTERPLOTS CONTINUED

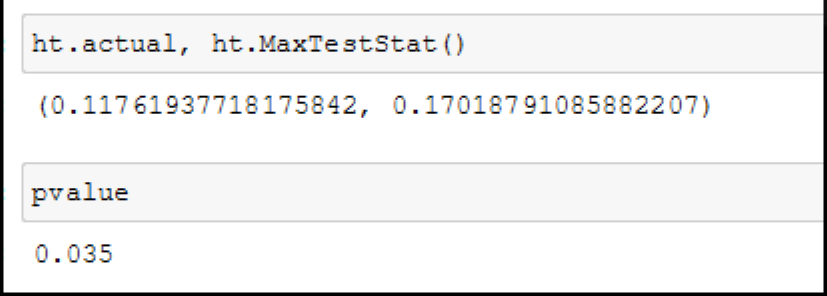
- Within the correlation heatmap, we can see and ascertain the following:
  - The strongest, positive correlation exists between packed_cell_volume and mucous_membrane.
  - capillary_refill_time and packed_cell_volume is the second strongest, positive correlation.
  - age and pulse are the strongest, negative correlation.
  - age and respiratory_rate is a moderately strong, negative correlation.

HYPOTHESIS TESTING

- For my hypothesis testing, I elected to conduct the "Testing a Correlation" hypothesis test - this is primarily because I am interested in the relationships of my variables. For my specific hypothesis, below, I will run this correlation test to determine whether I have support for my hypothesis or not.

- Hypothesis:
  - I hypothesize that horses of which 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.

HYPOTHESIS TESTING



```
ht.actual, ht.MaxTestStat()
```

```
(0.11761937718175842, 0.17018791085882207)
```

```
pvalue
```

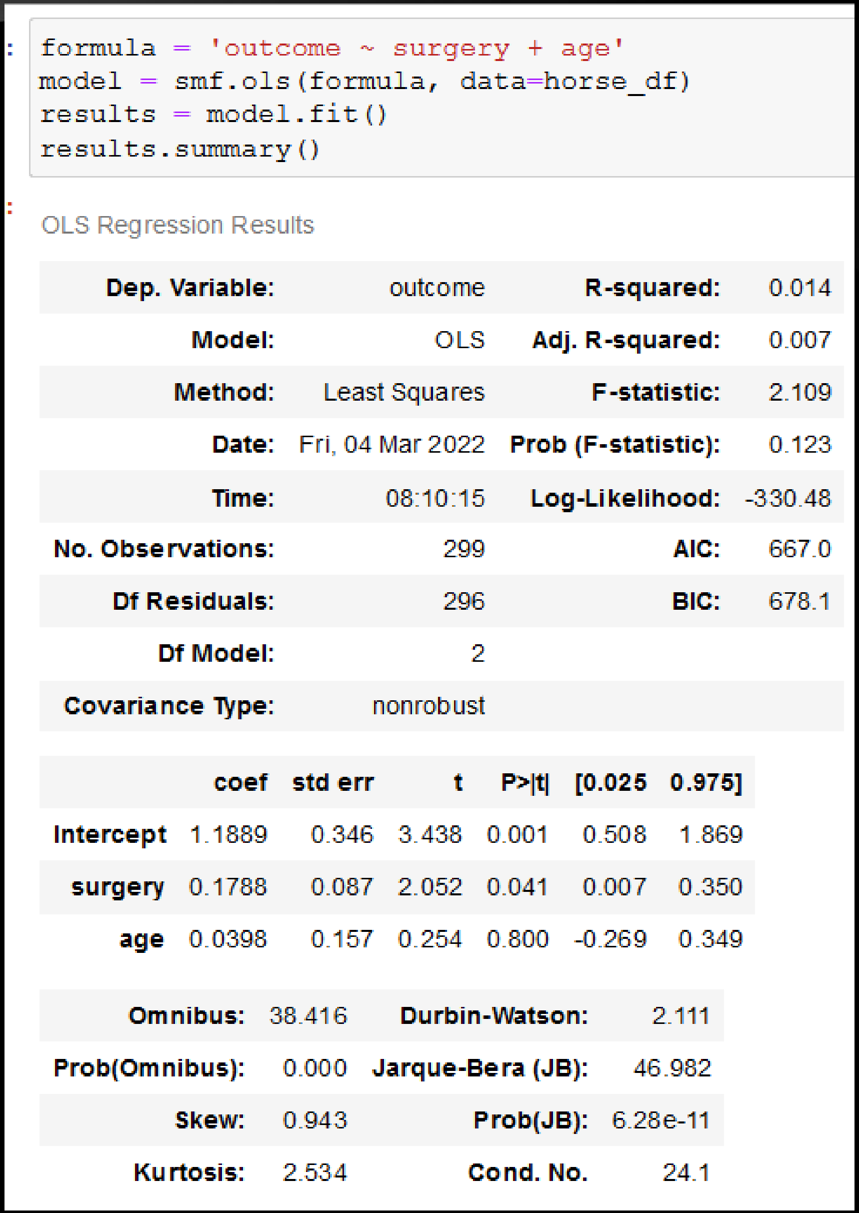
```
0.035
```

- The actual correlation for the relationship between surgery option and horse life outcome was a positive, weak relationship, coming in at 0.12 for the corellation coefficient. After 1,000 iterations, this value went up to 0.17 for the correlation coefficient. Overall, we can see that the observed correlation is small and weak.
- However, the computed p-value is 0.04 (which is less than 0.05), which indicates that this relationship is statistically significant.
- So, while the observed correlation is small, the p-value tells us that there is likely support for my hypothesis - particularly the second part of my hypothesis in which I hypothesize the relationship between these variables is statistically significant.

REGRESSION ANALYSIS

- For this section, I wanted to consider a few relationships, so I did a couple of extra analyses.
- Below is a summary of the definitions of the output table characteristics/stats that I will be looking at for the regressions I run. I will not be discussing all of the information, but rather those characteristics that are most important for my project scope and topic.
    - The dependent variable is the variable we are looking at to see how it is impacted by the predictor or independent variables.
    - Degrees of freedom are an accounting of how many parameters are estimated by the model and, by extension, a measure of complexity for linear regression models.
    - The df model indicates the number of predictor variables.
    - The covariance type for the below measures is nonrobust - this is a measure of how two variables are linked in a positive or negative way, and a robust covariance is one that is calculated in a way to minimize or eliminate variables, which is not the same as nonrobust, of course.
    - R-squared is probably the most important measurement here, as it is the measurement of how much of the dependent variable is explained by the changes in our independent variables.
    - Adjusted R-squared is important for analyzing multiple dependent variables' efficacy on the model.
    - The F-statistic is comparing your produced linear model for your variables against a model that replaces your variables' effect to 0, to find out if your group of variables are statistically significant. To interpret this number correctly, using a chosen alpha value and an F-table is necessary. Prob (F-Statistic) uses this number to tell you the accuracy of the null hypothesis, or whether it is accurate that your variables' effect is 0.

REGRESSION ANALYSIS CON'T.



```
formula = 'outcome ~ surgery + age'
model = smf.ols(formula, data=horse_df)
results = model.fit()
results.summary()
```

OLS Regression Results

Dep. Variable:	outcome	**R-squared:**	0.014
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	2.109
Date:	Fri, 04 Mar 2022	**Prob (F-statistic):**	0.123
Time:	08:10:15	**Log-Likelihood:**	-330.48
No. Observations:	299	**AIC:**	667.0
Df Residuals:	296	**BIC:**	678.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.1889	0.346	3.438	0.001	0.508	1.869
surgery	0.1788	0.087	2.052	0.041	0.007	0.350
age	0.0398	0.157	0.254	0.800	-0.269	0.349

Omnibus:	38.416	**Durbin-Watson:**	2.111
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	46.982
Skew:	0.943	**Prob(JB):**	6.28e-11
Kurtosis:	2.534	**Cond. No.**	24.1

REGRESSION ANALYSIS CON'T.

For the first relationship (previous slide), I wanted to look at the impact that surgery option plus the accompaniment of age had on a horse's life outcome. From the results above, we can see that (using the model of Ordinary Least Squares):

* The dependent variable is outcome (of course).

* There are 296 degrees of freedom.

* The df model is 2.

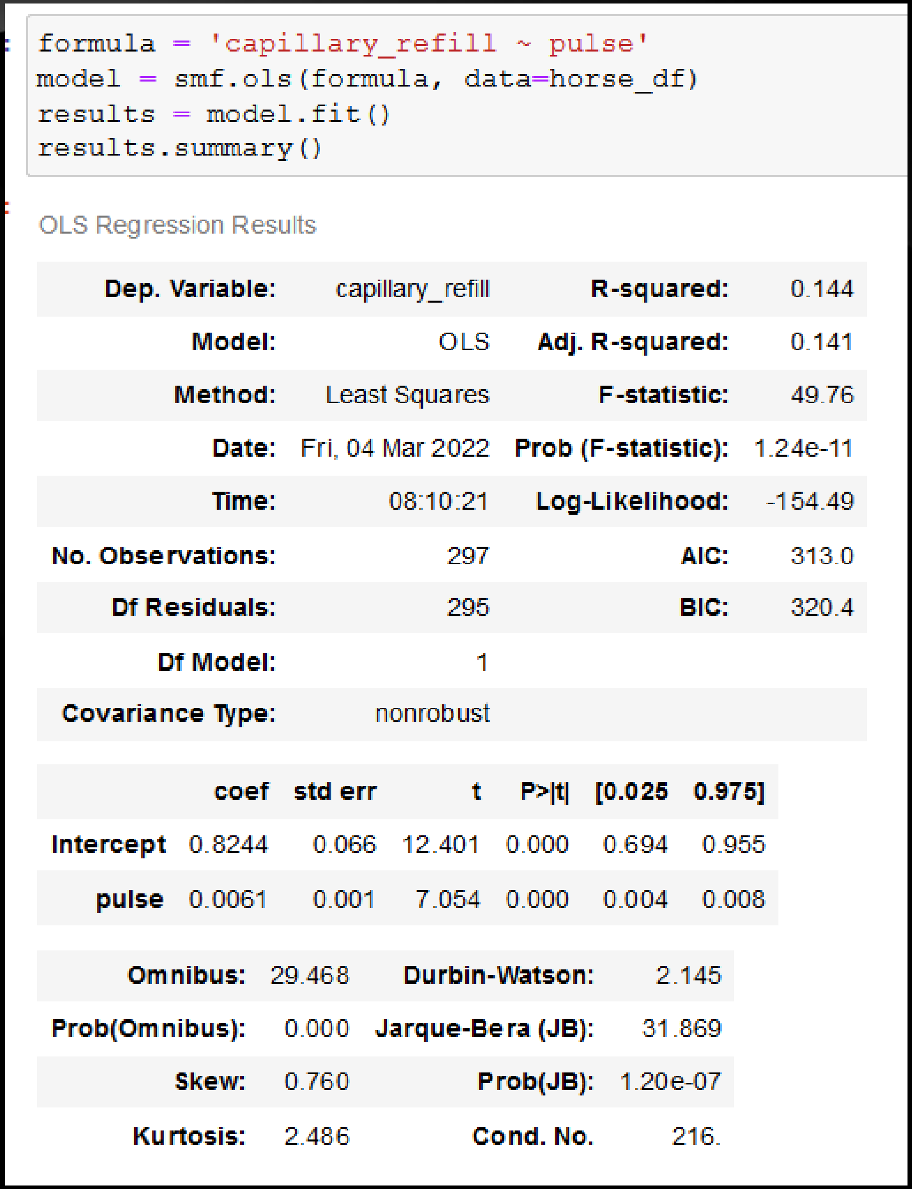
* The covariance type is nonrobust.

* R-squared is very low (0.014), which indicates that the model explains only 1.4% of the change in the outcome variable.

* Adjusted R-squared is also not very strong - 0.7%.

* Our F-1 score statistic is 2.109 with its probability being .000123%.

REGRESSION ANALYSIS CON'T.



```
formula = 'capillary_refill ~ pulse'
model = smf.ols(formula, data=horse_df)
results = model.fit()
results.summary()
```

OLS Regression Results

Dep. Variable:	capillary_refill	**R-squared:**	0.144
Model:	OLS	**Adj. R-squared:**	0.141
Method:	Least Squares	**F-statistic:**	49.76
Date:	Fri, 04 Mar 2022	**Prob (F-statistic):**	1.24e-11
Time:	08:10:21	**Log-Likelihood:**	-154.49
No. Observations:	297	**AIC:**	313.0
Df Residuals:	295	**BIC:**	320.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.8244	0.066	12.401	0.000	0.694	0.955
pulse	0.0061	0.001	7.054	0.000	0.004	0.008

Omnibus:	29.468	**Durbin-Watson:**	2.145
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31.869
Skew:	0.760	**Prob(JB):**	1.20e-07
Kurtosis:	2.486	**Cond. No.**	216.

REGRESSION ANALYSIS CON'T.

For the second relationship, I wanted to look at the relationship between a capillary refill time and pulse. From the results above, we can see that (using the model of Ordinary Least Squares):

* The dependent variable is capillary_refill.

* There are 295 degrees of freedom.

* The df model is 1.

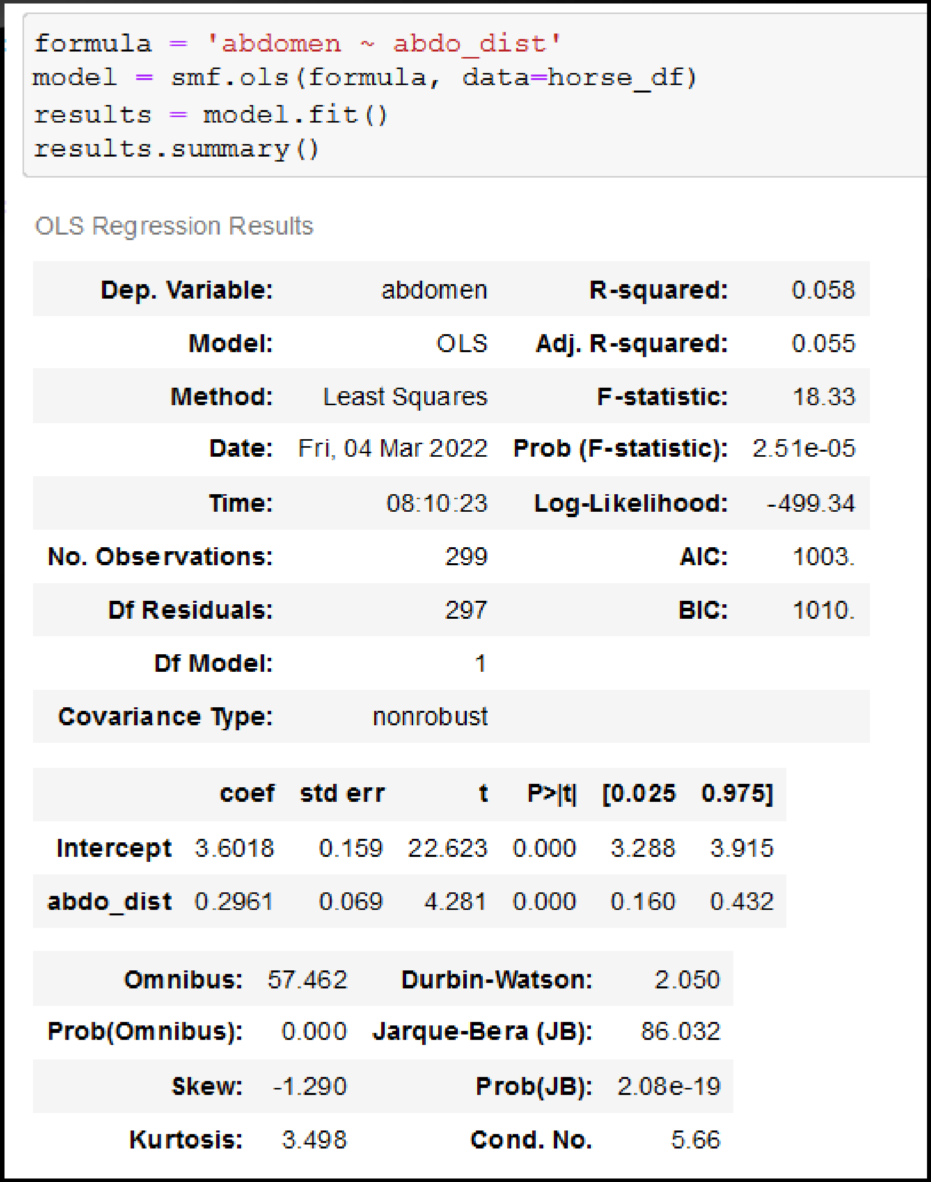
* The covariance type is nonrobust.

* R-squared is somewhat low (0.144), which indicates that the model explains only 14.4% of the change in the outcome variable.

* Adjusted R-squared is similarly strong - 14.1%.

* Our F-1 score statistic is 49.76 with its probability being .00124%.

REGRESSION ANALYSIS CON'T.



```
formula = 'abdomen ~ abdo_dist'
model = smf.ols(formula, data=horse_df)
results = model.fit()
results.summary()
```

OLS Regression Results

Dep. Variable:	abdomen	**R-squared:**	0.058
Model:	OLS	**Adj. R-squared:**	0.055
Method:	Least Squares	**F-statistic:**	18.33
Date:	Fri, 04 Mar 2022	**Prob (F-statistic):**	2.51e-05
Time:	08:10:23	**Log-Likelihood:**	-499.34
No. Observations:	299	**AIC:**	1003.
Df Residuals:	297	**BIC:**	1010.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	3.6018	0.159	22.623	0.000	3.288	3.915
abdo_dist	0.2961	0.069	4.281	0.000	0.160	0.432

Omnibus:	57.462	**Durbin-Watson:**	2.050
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	86.032
Skew:	-1.290	**Prob(JB):**	2.08e-19
Kurtosis:	3.498	**Cond. No.**	5.66

REGRESSION ANALYSIS CON'T.

For the third relationship, I wanted to look at the relationship between a horse's abdomen status and its abdominal distension. From the results above, we can see that (using the model of Ordinary Least Squares):

* The dependent variable is abdomen.
* There are 297 degrees of freedom.
* The df model is 1.
* The covariance type is nonrobust.
* R-squared is low (0.058), which indicates that the model explains only 5.8% of the change in the outcome variable.
* Adjusted R-squared is also not very strong - 5.5%.
* Our F-1 score statistic is 18.33 with its probability being .00251%.

CONCLUSION.

- Overall:
  - Additional future analyses may be helpful.
  - Having more datapoints (instead of just 299) may also prove helpful in the future to bolster the sample size.
  - 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.
  - For some of the variables, such as capillary_refill_time, it 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.
  - Moving forward, the above are the additional future considerations I have.