**Horse colic 이란 무엇인가?**

산통이란 수의학적인 용어이고 쉽게 말해서 배앓이를 말하는 것이다. 말은 큰 체구에 비해 위의 용적이 10리터 정도밖에 되지 않을 정도로 위가 적으며 말은 소와 달리 되새김을 할 수가 없어서 토해낼 수가 없는 특징을 갖고 있다 이러한 특이한 생리구조를 가졌기에 산통이 자주 발생하는데 대표적인 산통으로는 위식체, 소장감돈, 소장교액, 회맹중첩, 대결장변위 또는 염전, 대결장폐색 등 다양하고, 어떠한 원인에 의해서 발생하더라도 겉으로 드러나는 증상은 대개 비슷하다고 한다. 특히 산통이 심한 경우 수술을 하기도 하고 사망하거나 안락사를 시키기도 한다. 그렇다면 산통이 발생했을 경우 의학적 상태에 따라 말의 생존 여부를 예측 할 수 있을까?

분석에는 UCI에서 제공하는 ‘Horse Colic Data Set’을 이용하였다.

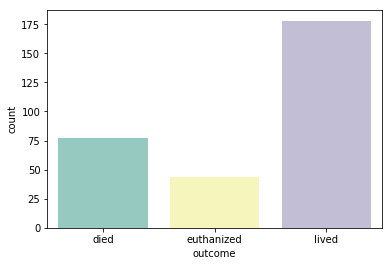
- 데이터 출저 : UC Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Horse+Colic>)

**변수 이름과 설명**

**산통이 발생했을 때 말의 생존여부는 어떻게 될까?**

먼저 산통이 발생 했을 때 말의 생존 여부를 확인해보면 약 60%정도의 말이 생존 하지만, 죽거나 안락사 한 비율을 합쳐, 사망한 말의 비율은 40%정도로 굉장히 높은 비율이다. 이것으로 보아. 산통은 말에게 매우 치명적인 상황인 것을 다시 한번 확인 할 수 있다.

|  |  |  |  |
| --- | --- | --- | --- |
| outcome | Lived | Died | Was euthanized |
| Count | 178 | 77 | 44 |
| percentage | 59.3% | 25.7% | 14.7% |



**<figure 1> outcome countplot**

**data탐색**

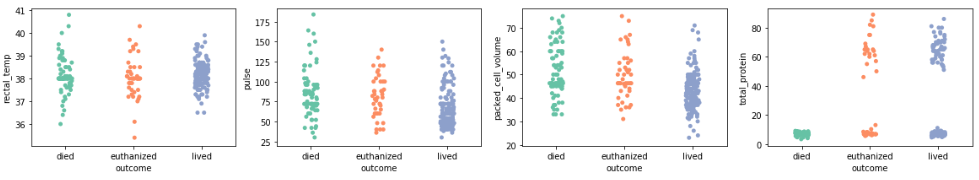
이제 데이터를 본격적으로 탐색하기 위하여 연속형 변수와 범주형 변수의 기초 통계량을 조사해 보았다. 연속형 변수의 기초 통계량 표는 아래와 같다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | | Value | | | | | | Missing |
| Mean | Median | Std Dev | Skewness | Kurtosis | range |
| **Age** | lived | 1.64 | 1 | 2.173972 | 3.080813 | 7.516494 | 1-9 | 0 |
| died |  |  |  |  |  |  |  |
| euthanized |  |  |  |  |  |  |  |
| **Rectal\_temperature** | lived | 38.162921 | 38.2 | 0.569344 | 0.036021 | 0.845986 | 36.5-39.9 | 26 |
| died | 38.12337 | 38.1 | 0.925776 | 0.338782 | 0.813765 | 36.0-40.8 | 24 |
| euthanized | 38.04090 | 38.05 | 1.016302 | -0.13436 | 0.631468 | 35.4-40.3 | 10 |
| **Pulse** | lived | 64.00000 | 54.0 | 24.544886 | 1.391464 | 1.315963 | 30.0-150.0 | 12 |
| died | 85.05194 | 86.0 | 31.730982 | 0.796242 | 0.793754 | 30.0-184.0 | 11 |
| euthanized | 81.52272 | 82.0 | 26.001001 | 0.174706 | -0.565632 | 36.0-140.0 | 1 |
| **respiratory rate** | lived | 29.48470 | 24.0 | 18.290945 | 1.683889 | 2.818225 | 8.0-96.0 | 31 |
| died | 32.39936 | 30.0 | 16.424164 | 1.960948 | 5.05072 | 12.0-96.0 | 19 |
| euthanized | 31.01556 | 24.0 | 16.965769 | 1.354574 | 1.323845 | 12.0-80.0 | 8 |
| **nasogastric reflux PH** | lived | 4.684619 | 5.15 | 1.995171 | -0.297241 | -1.343638 | 1.0-7.5 | 154 |
| died | 4.895197 | 5.35 | 1.638067 | -0.584413 | -0.231025 | 1.5-7.5 | 57 |
| euthanized | 4.471913 | 2.0 | 2.200063 | 0.481735 | -1.625962 | 1.0-7.0 | 35 |
| **packed cell volume** | lived | 42.98874 | 42.0 | 7.974895 | 0.724847 | 1.446866 | 23.0-71.0 | 13 |
| died | 51.88908 | 53.0 | 11.402502 | 0.113267 | -0.930693 | 33.0-75.0 | 8 |
| euthanized | 49.96498 | 50.5 | 11.361869 | 0.29834 | -0.643649 | 31.0-75.0 | 8 |
| **total protein** | lived | 25.76657 | 7.5 | 28.452084 | 0.736292 | -1.352775 | 4.5-86.0 | 13 |
| died | 6.829870 | 6.8 | 1.177661 | -0.612426 | 0.539978 | 3.3-9.0 | 12 |
| euthanized | 35.622727 | 56.0 | 30.705228 | -0.116664 | -1.760177 | 5.5-89.0 | 8 |
| **abdomcentesis total protein** | lived | 2.247753 | 2.0 | 1.924404 | 1.046526 | 0.266961 | 0.1-8.0 | 126 |
| died | 2.671429 | 3.6 | 2.13826 | 1.815788 | 3.446602 | 1.4-10.1 | 53 |
| euthanized | 2.209091 | 2.0 | 1.447043 | 2.128424 | 5.384482 | 1.0-7.4 | 19 |

**<Table 1> 연속형 변수의 기초통계량**

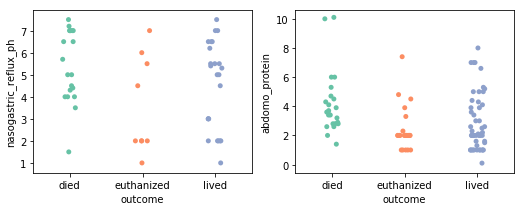
가장 먼저 눈에 띈 변수는 total protein이다. 죽은 경우를 보았을 때 평균이 6.8이고 편차는 매우 적었다. 그 외에는 평균은 25.7과 35.6이지만 편차가 매우 높은 것을 확인 할 수 있다. Rectal temperature는 살았을 때 온도 편차가 적고 다른 상황에서는 편차가 큰 것을 볼 수 있다. 정상 온도의 범위를 넘어섰을 때는 몸의 상태가 좋지 못한 경우라고 생각하니 생존여부를 예측하는데 좋은 변수가 될 듯 하다. Pulse와 packed cell volume는 평균과 분산을 비교해 봤을 때 생존 여부에 따라 차이가 나는 것을 확인 하였다.

변수들 사이에 어떤 차이가 있는지 시각적으로 확인하기 위하여 python seaborn 패키지의 stripplot 함수를 이용하여 그래프를 그려보았다.



그래프를 보았을 때 total protein의 분포 차이가 확연히 구분되며 Rectal temperature와 Pulse는 분포는 비슷하나 중앙 집중도의 차이가 보인다. 변수의 분포를 좀 더 자세히 분석하면 좋겟지만, 통계적으로 깊은 이해도를 요하는 문제로 보이므로 다음 기회에 자세히 분석하기로 하자.

그리고 기초통계량에서는 차이가 보이지만 missing value 비율이 상당히 높아 신뢰성이 떨어지는 변수로는 nasogastric reflux PH, abdomcentesis total 2가지 변수가 보인다.

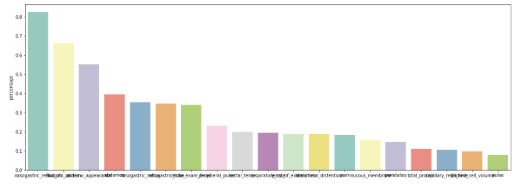


마찬가지로 그래프를 그려 시각화 하였지만 변수의 개수가 너무 적어 좋은 변수라고 보기는 어려울 것 같다.

범주형 데이터의

**Missing value**

앞서 기초 통계량 표에서 확인 할 수 있듯이 UCI의 ‘Horse Colic Data Set’의 Missing value 비율은 전체 데이터의 30%정도이며, 50 ~ 80%의 missing value를 가진 변수들도 존재한다. 이로 보아 missing value를 잘 대체할 수 있는 통계량을 설정하는 것이 좋은 분석을 하기 위한 포인트가 될 것이다. 하지만 이는 통계적으로 어려운 지식이 필요하니 우리는 간단하게 Missing value 비율이 50%이상인 변수는 제외하고 연속형 변수는 평균값을, 범주형 변수는 최빈값 으로 설정하고 분석을 계속해보기로 한다



**<figure 2> missing value percentage plot**