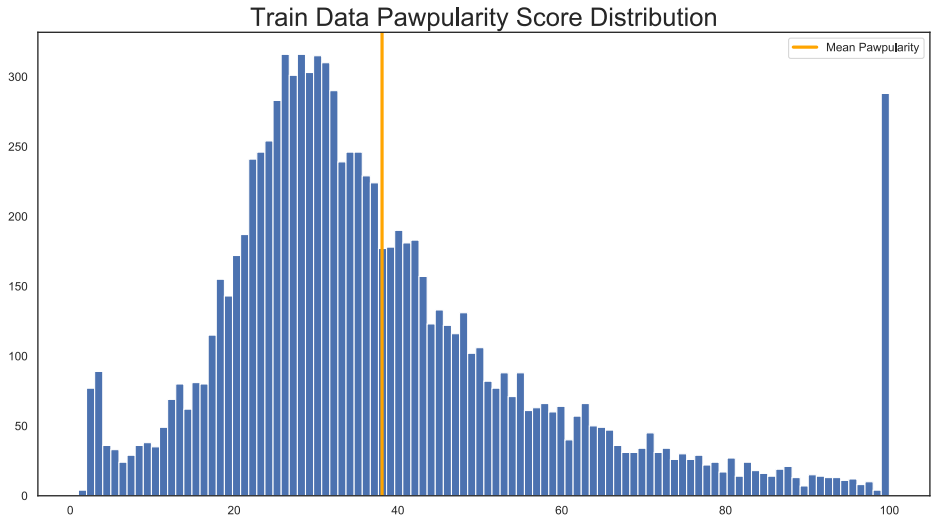
The dataset for this project has both image data and metadata that store some basic information about the image. The size of the metadata train.csv is 9912*14. It contains the ID and 12 basic features for each photo as well as the photo's Pawpularity score.

Following table shows some sample dataset. The ID column gives the photo's unique Pet Profile ID corresponding the photo's file name. The Pawpularity score are continuous number between 0 and 100. The 12 features (Focus, Eyes, Face, Near, Action, Accessory, Group, Collage, Human, Occlusion, Info and Blur) are labeled with the value of 1 (Yes) or 0 (No).

	Id	Subject Focus	Eyes	Face	Near	Action	Accessory	Group	Collage	Human	Occlusion	Info	Blur	Pawpularity
0	0007de18844b0dbbb5e1f607da0606e0	0	1	1	1	0	0	1	0	0	0	0	0	63
1	0009c66b9439883ba2750fb825e1d7db	0	1	1	0	0	0	0	0	0	0	0	0	42
2	0013fd999caf9a3efe1352ca1b0d937e	0	1	1	1	0	0	0	0	1	1	0	0	28
3	0018df346ac9c1d8413cfcc888ca8246	0	1	1	1	0	0	0	0	0	0	0	0	15
4	001dc955e10590d3ca4673f034feeef2	0	0	0	1	0	0	1	0	0	0	0	0	72

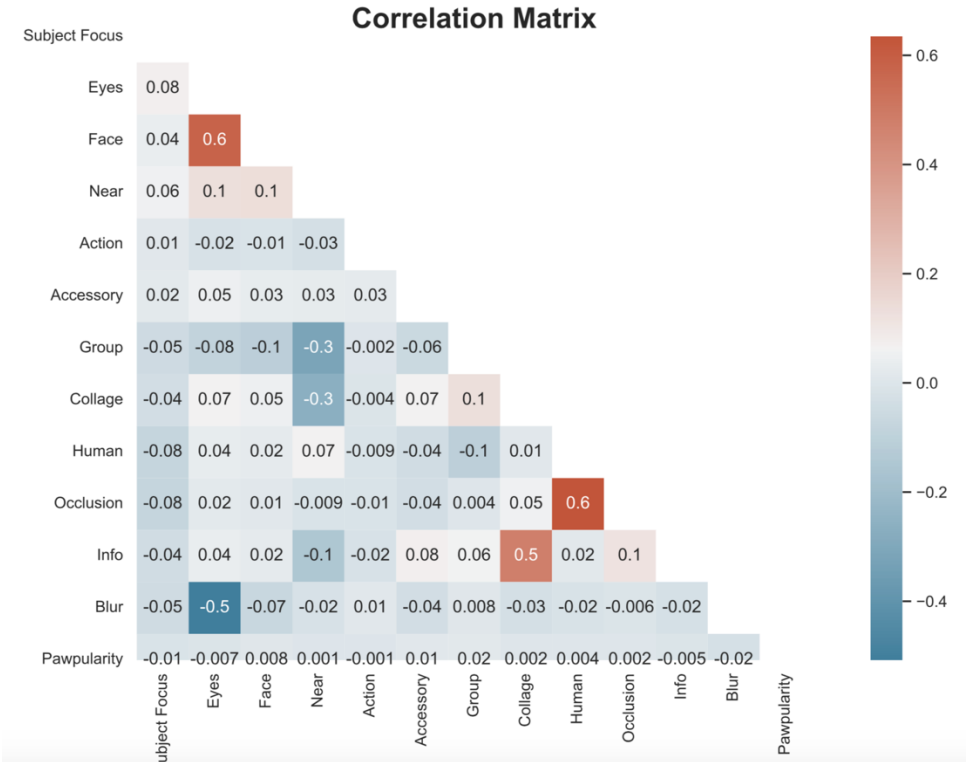
There are three interesting findings in train.csv.

1. Pawpularity Score Distribution



We can see that the distribution of pawpularity score is skewed. There is a small curve close to zero Pawpularity and close to 300 pawpularity scores at 100. The score is centered around 38 and has a peak on 0 and 100. Most of the photos don’t have high scores.

2. Correlations between predictor variables and responsor



There exists two high correlation (>0.5) between two pairs of the predictors. One is Occlusion and Human (Humans can hide part of the animal), another is Face and Eyes. Then I used VIF to check if there any multicollinearity exists in the data. Here is the VIF for each feature.

	feature	VIF
1	Eyes	2.307956
8	Human	1.723615
9	Occlusion	1.721186
2	Face	1.712161
11	Blur	1.510249
7	Collage	1.392512
10	Info	1.328935
3	Near	1.213886
6	Group	1.142908
0	Subject Focus	1.019446
5	Accessory	1.018282
4	Action	1.002690

Since all of the VIF are less than 2.5, there is no multicollinearity in the data.

3. No correlation between the predictors and responsor.

By looking at the last row of correlation matrix, the correlation between predictors and responsory is so small. This might mean that a good solution to this problem will require using the images but not the .csv metadata.