Action Report

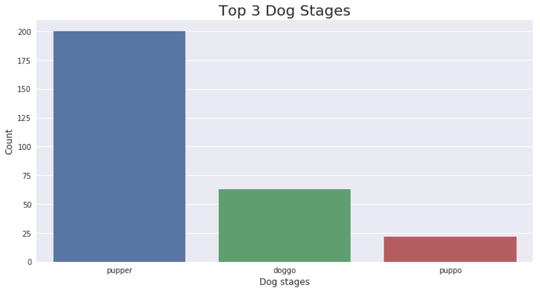
Created By: Ayobami Yusuf

Introduction:

After gathering, assessing, and cleaning the data, I then went ahead to deep-dive into the data to uncover some insights and presented these in insights in stunning and compelling visualizations which makes it easier to understand what's going on with the data.

The first question I investigated was "How popular are the different dog stages tweeted about?" and to answer this question, I wrote a program that counts the number of occurence of each dog stage and sorted in descending to identify the top 3 most popular dog stages. The code that produced the result and the output chart are given below:

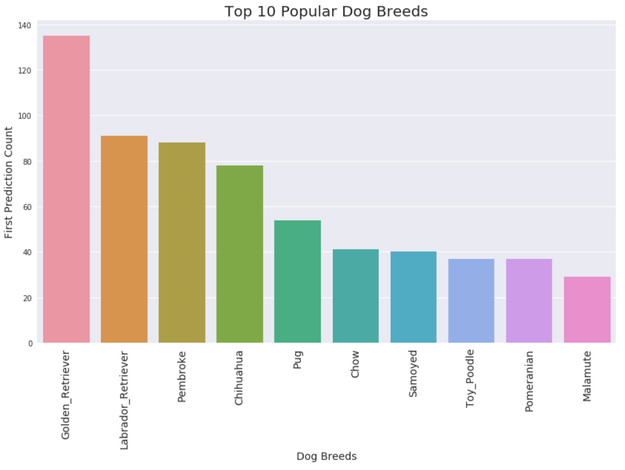
In [6]:
```
plt.figure(figsize = (12,6))
sns.set(style = 'darkgrid')
sorted_dog_stages = merged_df['stage'].value_counts().head(3).index
sns.countplot(data = merged_df, x ='stage', order = sorted_dog_stages, orient='v')
plt.xlabel('Dog stages', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.title('Top 3 Dog Stages', fontsize=20);
```



The second aspect of the data which I investigated was to find out the Top 10 Dog Breeds by First Predictions.

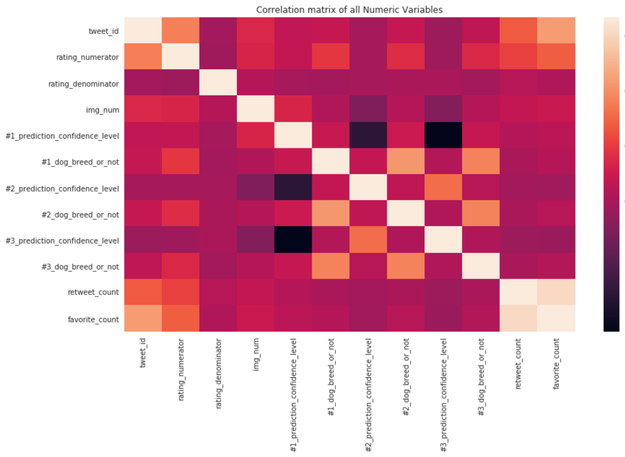
This required that I get the number of instances of each dog breed in the first predictions column, sort the output in descending order and subset the output to isolate only the top 10. My action is shown below:

In [7]:
```
plt.figure(figsize = (14,8))
plot = sns.barplot(x = merged_df['#1_prediction'].value_counts()[0:10].index,
            y = merged_df['#1_prediction'].value_counts()[0:10],
            data = merged_df);
plot.set_xticklabels(plot.get_xticklabels(),rotation = 90, fontsize = 14);
plt.xlabel("Dog Breeds",fontsize = 14);
plt.ylabel("First Prediction Count",fontsize = 14);
plt.title("Top 10 Popular Dog Breeds",fontsize = 20);
```



I then proceeded to investigate if any correlation exist between any pair of variables in the dataframe. This revealed interesting insights as shown below:

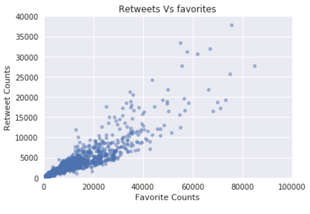
In [8]:
```
plot = plt.figure(figsize=(14,8))
sns.set_style('darkgrid')

correlations = merged_df.corr()
sns.heatmap(correlations,
            xticklabels=correlations.columns.values,
            yticklabels=correlations.columns.values);
plt.title('Correlation matrix of all Numeric Variables');
```



Seeing that only favorite_count and retweet_count showed some sort of interesting correlation, I zoomed in on those pair of variables to inspect their correlation in isolation. This is shown below:

In [9]:
```
sns.set_style('darkgrid')
plt.figure(figsize=(16,10))
merged_df.plot(kind = 'scatter', x = 'favorite_count', y = 'retweet_count', alpha = 0.5)
plt.xlim((0,100000))
plt.ylim((0,40000))
plt.xlabel('Favorite Counts')
plt.ylabel('Retweet Counts')
plt.title('Retweets Vs favorites');
```

```
<matplotlib.figure.Figure at 0x7f3c3399eb70>
```



From all the above, some of the insights gleaned include:

1. **It appears that WeRateDogs followers are more likely to retweet dog ratings' posts that they click the favorite button on as a strong positive correlation exist between retweet_count and favorite_count.**
2. **Dogs at the 'pupper' stage are the most featured on WeRateDogs Twitter ratings posts. So, one is likely more engagements on puppers, more than dogs at other stages**
3. **Apparently, that a dog has a higher rating does not guarantee higher posts engagements as ratings have low correlation with both retweet_count and favorite_count**
4. **The Golden Retriever is the most popular dog breed, followed by Labrador Retriever and Pembroke in Top 3**