We decided to make a different model for cats and different model for dogs due to the drastically different outcomes. While they have similar adoption outcomes, they have very different outcomes and there are many intrinsic and societal differences between outcomes between the two

Features to Clean / Feature Engineering

- **ID**
  - Get rid this attribute ✅
    - Make sure there are no duplicates after dropping the ID
- ~~**Name**~~
  - ~~**Cleaning**~~
    - ~~Maybe we should drop bc test.csv doesn't have name ☠️✅~~
    - ~~If there are numbers~~
      - ~~Change to Missing Name, BUT check~~
        - ~~If it was part of the name, keep as OG name~~
      - ~~Some are just the ID~~
      - ~~Some are just~~
  - ~~**Engineering**~~
    - ~~We are going to put names in 3 categories: Missing Name, Named at Shelter, Name Outside of Shelter~~
      - ~~This might be changed to just 2 categories if we use different models~~
- **Intake Time**
  - **Cleaning**
    -
  - **Engineering**
    - Use intake time and transform into intake hour, intake month, intake year ✅
    - Creates an (optional) function where we bucket hours into times of day (morning / afternoon / evening) and bucket months into seasons (winter, spring, summer, fall) ✅
    - When models need time as a *numerical* interpretation, we use sin and cos for hour and month in order for the model to understand that hour and month are cyclical ✅
- **Found Location**
  - **Engineering**
    - ~~Encode with longitude and latitude columns? - run it through an Ai~~

        - ~~Make a Found_A_Corner column~~
            - Realized outcome types based on if they were found on a corner were the same as others

- **Intake Type**
    - **Cleaning**
        - Drop the record found from wildlife ✅
    - **Engineering**
        - One hot encode the different intake types

- **Intake Condition**
    - **Cleaning**
        - Combine the Other column with Unknown (“Other / Unknown) ✅
        - ~~Drop ID A837091 (copy of a previous one) and~~ move ‘Space’ into Other / Unknown ✅
        - Put Behavior with Normal → they match the overall distribution in outcome types ✅
        - Combine Nursing / Neonatal → match each other’s distributions and are similar ✅
        - Combine Neurological with Med Urgent ✅
        - Combine Congenital with Sick ✅
        - Group these within Medical
            - Cats
                - Put Agonal with Med Urgent ✅
            - Dogs
                - Combine Parvo with Med Urgent ✅
        - Things we are LEAVING as is + explanations
            - Feral - did NOT match outake distributions and were way less likely to get adopted
            - Aged - did not match cat / dog distr.
            - CAT
                - Injured (more euthansia)
                - Med Attn + Med Urgent should remain the same (higher transfer rates
            - Dogs
                - Med Attn should be on its own
                - Med Urgent should be on your own
            - Injured, sick - dists did NOT match medical / outtake dists
    - **Engineering**
        - One hot encoding

- **Animal Type**
  - **Feature Engineering**
    - ~~This is the most PIVOTAL factor - we should either split initially, have this be the most important part~~
    - We decided to NOT split on animal type anymore and keep it all in one model

- **Sex upon Intake**
  - **Cleaning**
    - Missing values will be marked with “Unknown” ✅
  - **Engineering**
    - One hot encode yippee

- **Age upon Intake**
  - **Cleaning**
    - Drop the one value where age on intake is missing – it’s a test value ✅
    - Negative Numbers can be changed to 0 ✅
    - Converting everything into decimal - age in terms of ~~YEARS~~ MONTHS

- **Breed**

```
Crosstab (Cats, Top 10 Breeds vs Outcome) as Row Percentages:
outcome_type     Adoption  Died  Euthanasia  Return to Owner  Transfer
is_top_10_breed
0                   52.77  1.37        4.44             7.62     33.81
1                   49.74  1.49        3.78             3.91     41.07

Crosstab (Dogs, Top 10 Breeds vs Outcome) as Row Percentages:
outcome_type     Adoption  Died  Euthanasia  Return to Owner  Transfer
is_top_10_breed
0                   49.23  0.47        2.01            23.60     24.70
1                   48.93  0.54        3.35            23.63     23.54

Crosstab (Cats, is_mix vs Outcome) as Row Percentages:
outcome_type  Adoption  Died  Euthanasia  Return to Owner  Transfer
is_mix
0                56.83  1.66        3.05             3.63     34.82
1                45.90  1.37        4.31             4.60     43.81

Crosstab (Dogs, is_mix vs Outcome) as Row Percentages:
outcome_type  Adoption  Died  Euthanasia  Return to Owner  Transfer
is_mix
0                49.08  0.52        2.03            23.19     25.17
1                49.14  0.48        2.78            23.85     23.74
```

  - **Cleaning**

        - Breeds were redone in order to match the DogBreed Name in dogsbreed.csv
        - Cats
            - Top 10 breeds are adopted more, transferred less, and return to owner more
            - Cats that are not mixes are adopted 11% more and transferred 9% less
        - Dogs
            - Top 10 breeds vs not top 10 are pretty much the same
            - Is mix vs not mix are pretty much the same

    - **Engineering**
        - Dogs
            - Idk if mis mix or top 10 breeds would be helpful here but it won't hurt
        - Cats
            - Definitely should do is mix and top 10 breed since there is a clear separation between the two groups
        - Breed popularity: create new column “breed popularity”. A record will have a breed popularity equal to how frequently its breed shows up in the dataset. We then normalize these values.
        - Merged with a **dog_breeds.csv** dataset to get more features
            - Dog Breed Groups
        - Clustering dog breeds to get some dog breed groups
            - Genetic Lineage Clustering -- group breeds into clusters; Cluster breeds based on similarity in traits and assign a breed_cluster label

- **Color**
    - **Cleaning**
        - Combine cat colors from conventional knowledge
            - Blue tabby + silver tabby + gray + gray tabby + silver
            - Orange tabby + orange + orange tiger + red + red tabby + red tick + yellow
            - Cream + tan
            - Calico + tricolor
    - **Engineering**
        - Create a “primary_color” column that’s just the first part of any / in the colors ?

        - "Top_10_color" says whether this cat has one of the top 10 colors (is common)

- **Outcome Time?**
    - **Cleaning**
        - Ensure out year is AFTER intake year - correct the ones that aren't
            - Intake_ - 2/24/2019
            - Outtake - 7 /26/2019

- **Date of Birth**
    - Drop

- **Size -** gotten from merging with dogscleaned

This is for **EVERYTHING**

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
size					
1	0.445882	0.006358	0.019952	0.231674	0.296134
2	0.493356	0.013665	0.036086	0.066598	0.390294
3	0.500668	0.004421	0.033851	0.246857	0.214203
4	0.532284	0.004744	0.022044	0.219209	0.221720
5	0.492578	0.005848	0.017544	0.195682	0.288349

This is for specifically the **DOGS**

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
size					
1	0.445882	0.006358	0.019952	0.231674	0.296134
2	0.435766	0.003940	0.016444	0.273381	0.270469
3	0.500668	0.004421	0.033851	0.246857	0.214203
4	0.532284	0.004744	0.022044	0.219209	0.221720
5	0.492578	0.005848	0.017544	0.195682	0.288349

- Testing out dog_breeds.csv attributes - useful ones (missing 15K entries 🤦):
  - Good for Novice Owners

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Good For Novice Owners					
1	0.540397	0.004233	0.023464	0.222908	0.208999
2	0.494618	0.004814	0.034889	0.249743	0.215936
3	0.502030	0.005316	0.019638	0.224659	0.248357
4	0.443862	0.006534	0.023284	0.244432	0.281888
5	0.279904	0.004785	0.020164	0.330144	0.365003
No	0.529412	0.000000	0.058824	0.235294	0.176471

  - Tolerates being Alone

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Tolerates Being Alone					
1	0.452652	0.005436	0.034069	0.262719	0.245124
2	0.531544	0.004861	0.020087	0.218431	0.225077
3	0.465330	0.004953	0.018632	0.234552	0.276533
4	0.357576	0.008081	0.012121	0.262626	0.359596
5	0.357060	0.007001	0.031505	0.316219	0.288215
Moderate	0.529412	0.000000	0.058824	0.235294	0.176471

  - **Kid-Friendly** - Dogs with low friendliness are more likely to have euthanasia

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Kid-Friendly					
1.0	0.352941	0.003676	0.069853	0.323529	0.250000
2.0	0.464189	0.005807	0.014712	0.221448	0.293844
3.0	0.514406	0.003458	0.018632	0.218978	0.244526
4.0	0.439645	0.005043	0.023572	0.260827	0.270913
5.0	0.489615	0.005443	0.028407	0.242370	0.234165

- Easy to Groom

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Easy To Groom					
1.0	0.319863	0.004974	0.027666	0.321107	0.326391
2.0	0.431605	0.006737	0.017150	0.287056	0.257452
3.0	0.495461	0.003782	0.019667	0.228190	0.252900
4.0	0.518325	0.004897	0.020379	0.211690	0.244708
5.0	0.495098	0.005210	0.029077	0.236568	0.234047

- General Health

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
General Health					
1.0	0.368814	0.004068	0.021695	0.303729	0.301695
2.0	0.455210	0.006240	0.026035	0.244209	0.268307
3.0	0.504938	0.004777	0.019560	0.216900	0.253825
4.0	0.476467	0.005014	0.032767	0.265763	0.219988
5.0	0.573250	0.004251	0.019621	0.184107	0.218770

- Energy Level

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Energy Level					
2.0	0.252381	0.003401	0.031293	0.334694	0.378231
3.0	0.452609	0.006390	0.022098	0.237487	0.281416
4.0	0.460307	0.004330	0.038657	0.263696	0.233009
5.0	0.521187	0.005311	0.019908	0.226384	0.227210

- Exercise Needs

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Exercise Needs					
1.0	0.467018	0.007622	0.024480	0.224274	0.276605
2.0	0.230946	0.002292	0.026361	0.353582	0.386819
3.0	0.404768	0.005161	0.021381	0.250922	0.317768
4.0	0.453154	0.005061	0.039224	0.271013	0.231548
5.0	0.534860	0.004831	0.018759	0.219668	0.221882

- Potential for Playfulness

outcome_type	Adoption	Died	Euthanasia	Return to Owner	Transfer
Potential For Playfulness					
1.0	0.352071	0.008876	0.053254	0.292899	0.292899
2.0	0.548387	0.001112	0.022247	0.243604	0.184650
3.0	0.489583	0.004167	0.015972	0.204861	0.285417
4.0	0.449239	0.005852	0.022208	0.235970	0.286732
5.0	0.493228	0.005052	0.027973	0.246225	0.227522

-