Springboard Data Analytics Course

January 2019 Cohort

Capstone I Project

05Feb2019

John Parsons

Project Summary

The objectives of this paper are to define the parameters of the first capstone project for the Data Analytics Course at Springboard. This paper will discuss the following; what is the problem to be solved for the PetAdoption Kaggle Challenge, who is the client in this study and why do they care about the study, what is the type of data being used for this study, how will this study reduce adoption time for animals in shelters and minimize euthanizations and how will the results be presented to the client.

What is the problem you want to solve?

The **PetFinder** Data set is a current Kaggle competition that will be used to pick the speed at which a pet is adopted based on 24 attributes of varying datatypes. The algorithms generated from the competition will be used to guide shelters and rescuers find homes for abandoned pets and minimize the number of euthanizations for pets that cannot be adopted out of the shelter.

The main business problem this study will focus on is to minimize shelter cost by reducing the time these pets stay in the shelters through machine learning algorithms. The cost varies between shelters, but it is around 20 dollars per day to shelter a pet and 200 dollars for the initial veterinarian bill (deworming, vaccinations, neutering and spaying) in the United States. The Histogram generated in R for Figure 1 shows the distribution of animals over five ordinal target variables. There are only 410 animals that were adopted on the same day for target value of 0 and 3,090 animals were adopted out between 8 and 30 days. The biggest concern is the 4,197 animals that stayed in the shelters for more than 100 days for the target value of 4. Every day these animals are living in the shelters for more than 100 days will cost 83,000 per day. The

goal is to significantly reduce the time these animals stay in the shelters and adopt them out before the 100-day period has lapsed for each animal.

The second problem is to enhance favorable publicity ratings for the Nestle Company which has purchased the PetFinder App. Nestles model is "pets and people are better together" and by finding those models that will successfully adopt out pets and minimize the number of euthanizations for unadoptable pets, is a win-win solution for the pets and the shelters.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

PetFinder was founded by Betsy and Jared Saul in 1996 in Pittstown, NJ. This became the largest online Pet Adoption company that has listed over 350,000 adoptees from 14,000 shelters across the world. PetFinder was purchased by the Nestle Purina Pet Care Company in June of 2013 and was the first major acquisition of a digital company. The champion model that is generated to decrease shelter time for pets will benefit the board of directors, investors, 18,000 employees and 400 scientists who work for the Nestle Petcare company to continue achieving their goals. The last major benefit is finding homes for over 250,000 pets that have yet to be adopted and make this a successful campaign for people and pets.

What data are you using? How will you acquire the data?

The PetFinder.my Adoption Dataset is coming from the Kaggle competition and it contains a total of three files to be used for the **AdoptionSpeed** at which an animal is adopted from the shelter (Figure 2). The dataset contains three files for determining the rate at which an animal is adopted from the shelter. The first file is a metadata set that contains a total of 24

attributes and 14,993 rows of data. There are no missing variables in the dataset as shown in Figure 3 from the **AmeliaView** in R, but Python using the ebola statement shows missing values for the Name and Description Attributes. The summary, description and missing data results from R can be seen in the appendix.

The **PetAdoption** Speed attribute will be the Target variable and has a total of five levels or categories. The remaining 23 attributes are the Input variables to be used to develop a model to predict adoption speed of a pet. Figure 7 shows the output in Python of the datatypes for all 24 attributes in the dataset. Almost all attributes are categorized as an Int64 except for the **Name, RescuerID, Description** and **PetID** which are an Object and **PhotoAmt** which is a Float64. The **PetId** and **RescuerID** are unique identifiers and will not be used in the analysis and the name and description will be used in the Text Analysis section of the study.

The second two files are unsupervised files and contains a set of pictures (jpg files) and text files (Jason files). These two files are an addition to the original training file and can be used to improve model's performance. The goal of this study is to use the training dataset to develop the best champion model to reduce shelter time for the pets based on the most significant attributes. The text files and photos will be analyzed separately to do sentiment analysis on the most favorable words used in this adoption campaign and which photos were most successful in adopting out these pets.

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

Python Notebook and SQL will be used for data preparation, cleaning and exploratory data analysis of the metadata set. Data wrangling includes smoothing outliers, removing redundant

and irrelevant input variables, minimizing the skewness and standard deviation from the attributes. This data set only contains missing values in the name and description of the set and the other 22 attributes have no missing values.

Data analysts need to reduce the number of attributes used for the models to reduce dimension space and help model performance. The p values, logworth or other factors can be used to determine the best attributes to use in the selected models. These values were initially determined by JMP SAS and the results can be seen in Figure 8. A total of eight attributes have been selected to be used as the input variables for the machine learning algorithms to predict adoption speed. The other attributes will be used to determine if they can increase the performance of the model or if they have no effect on the model's performance.

The first eight input variables will be used to develop several models (Decision Trees, Logistic Regression, Neural Nets, Random Forests and Ensemble Models) to determine the best approach for reducing Shelter time for these pets. The data will be divided into a training and validation data set (65:35 split). The validation data set will be used to determine the model's performance and the model that has the lowest misclassification rates will be selected as the champion model.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The deliverables from this project will contain the PetAdoption Final Report that will contain the following items; Executive Summary, Introduction, Data Set Description, Project Objectives, Data Preparation, Data Analysis Methods, Results and Future Recommendations based on this study. The Python code and SQL will be submitted in a text and or HTML File for this project. All files will be uploaded to GitHub to present all findings from this study

Appendix

Figure 1: Histogram of Pet Adoption Speed.

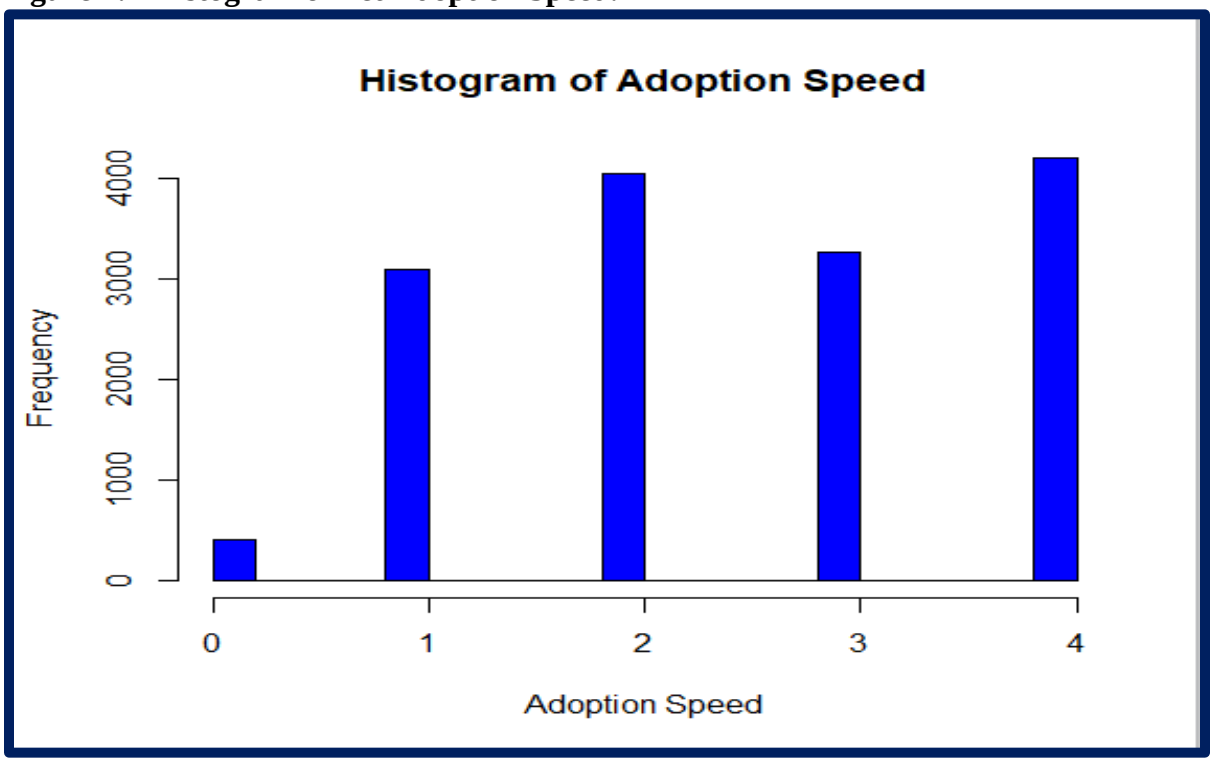


Figure 2: Snapshot of the datafiles for the PetFinder.my Adoption Prediction Challenge.

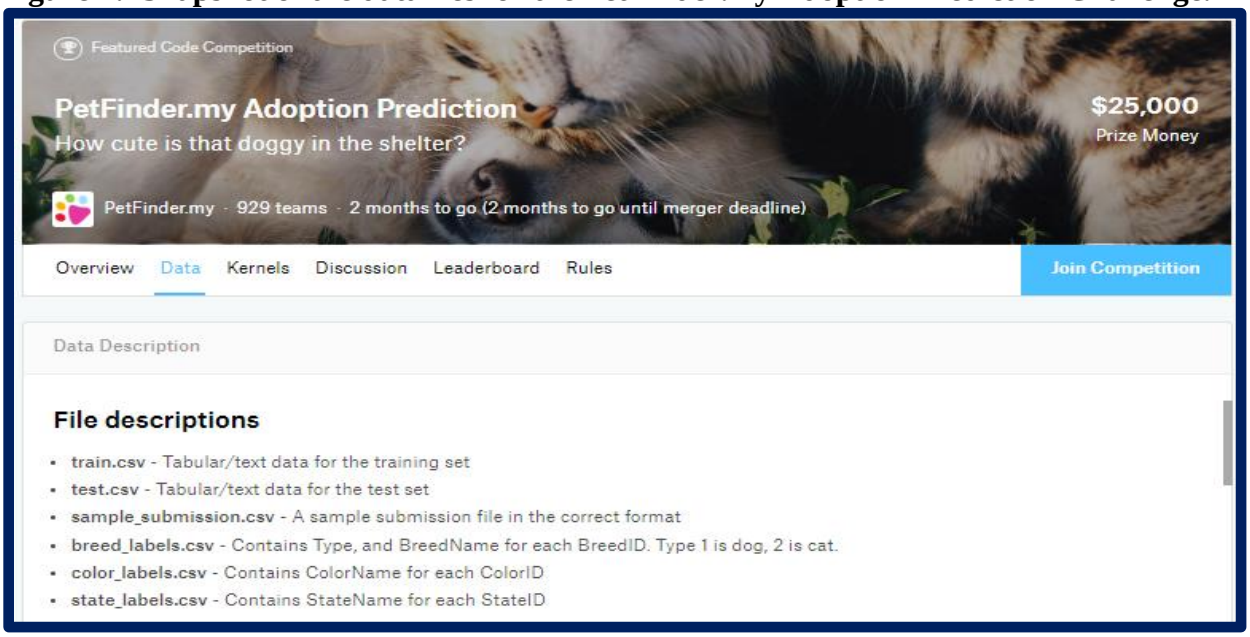
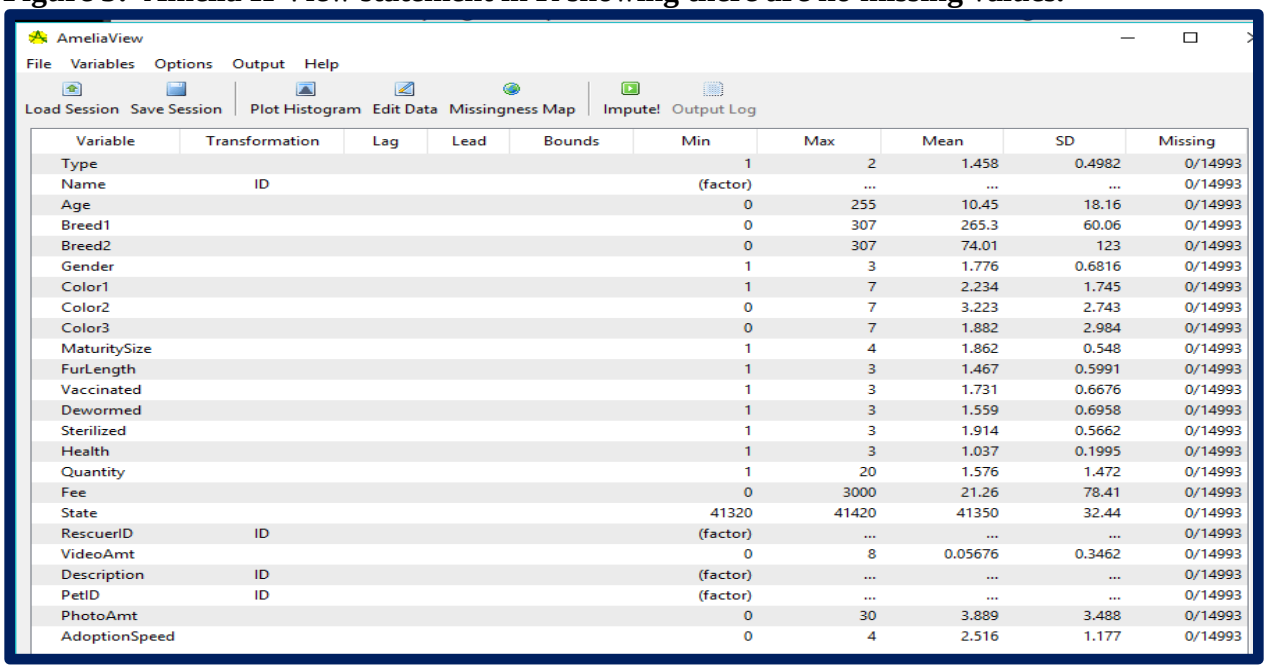


Figure 3: Amelia II View statement in R showing there are no missing values.



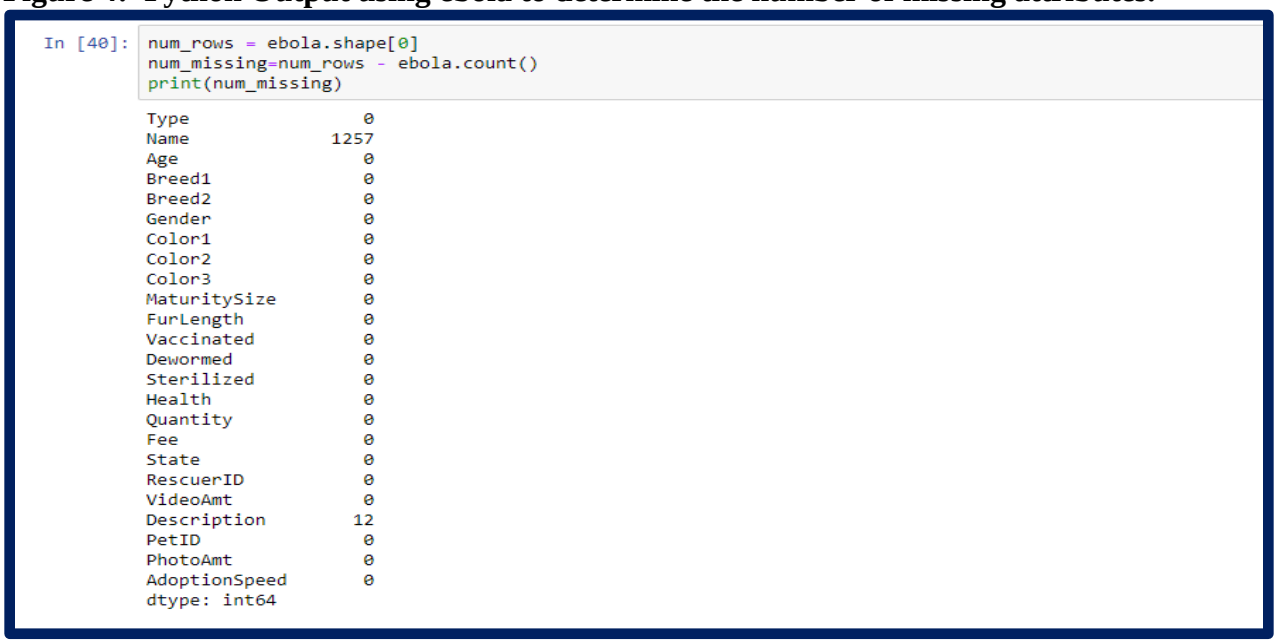
AmeliaView

File Variables Options Output Help

Load Session Save Session | Plot Histogram Edit Data Missingness Map | Impute! Output Log

Variable	Transformation	Lag	Lead	Bounds	Min	Max	Mean	SD	Missing
Type					1	2	1.458	0.4982	0/14993
Name	ID				(factor)	...	...	...	0/14993
Age					0	255	10.45	18.16	0/14993
Breed1					0	307	265.3	60.06	0/14993
Breed2					0	307	74.01	123	0/14993
Gender					1	3	1.776	0.6816	0/14993
Color1					1	7	2.234	1.745	0/14993
Color2					0	7	3.223	2.743	0/14993
Color3					0	7	1.882	2.984	0/14993
MaturitySize					1	4	1.862	0.548	0/14993
FurLength					1	3	1.467	0.5991	0/14993
Vaccinated					1	3	1.731	0.6676	0/14993
Dewormed					1	3	1.559	0.6958	0/14993
Sterilized					1	3	1.914	0.5662	0/14993
Health					1	3	1.037	0.1995	0/14993
Quantity					1	20	1.576	1.472	0/14993
Fee					0	3000	21.26	78.41	0/14993
State					41320	41420	41350	32.44	0/14993
RescuerID	ID				(factor)	...	...	...	0/14993
VideoAmt					0	8	0.05676	0.3462	0/14993
Description	ID				(factor)	...	...	...	0/14993
PetID	ID				(factor)	...	...	...	0/14993
PhotoAmt					0	30	3.889	3.488	0/14993
AdoptionSpeed					0	4	2.516	1.177	0/14993

Figure 4: Python Output using ebola to determine the number of missing attributes.



```
In [40]: num_rows = ebola.shape[0]
         num_missing=num_rows - ebola.count()
         print(num_missing)
```

```
Type              0
Name           1257
Age               0
Breed1            0
Breed2            0
Gender            0
Color1            0
Color2            0
Color3            0
MaturitySize      0
FurLength         0
Vaccinated        0
Dewormed          0
Sterilized        0
Health            0
Quantity          0
Fee               0
State             0
RescuerID         0
VideoAmt          0
Description      12
PetID             0
PhotoAmt          0
AdoptionSpeed     0
dtype: int64
```

Figure 5: Summary statement in R Studio from the metadata set.

```
> summary(mydata)
      Type            Name           Age             Breed1           Breed2           Gender      
 Min.   :1.000          : 1257   Min.   :  0.00   Min.   :  0.0   Min.   :  0.00   Min.   :1.000  
 1st Qu.:1.000   Baby   :   66   1st Qu.:  2.00   1st Qu.:265.0   1st Qu.:  0.00   1st Qu.:1.000  
 Median :1.000   Lucky  :   64   Median :  3.00   Median :266.0   Median :  0.00   Median :2.000  
 Mean   :1.458   Brownie:   54   Mean   : 10.45   Mean   :265.3   Mean   : 74.01   Mean   :1.776  
 3rd Qu.:2.000   No Name:   54   3rd Qu.: 12.00   3rd Qu.:307.0   3rd Qu.:179.00   3rd Qu.:2.000  
 Max.   :2.000   Mimi   :   52   Max.   :255.00   Max.   :307.0   Max.   :307.00   Max.   :3.000  
                 (Other):13446                                                                    
     Color1          Color2          Color3       MaturitySize     FurLength       Vaccinated   
 Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
 Median :2.000   Median :2.000   Median :0.000   Median :2.000   Median :1.000   Median :2.000  
 Mean   :2.234   Mean   :3.223   Mean   :1.882   Mean   :1.862   Mean   :1.467   Mean   :1.731  
 3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :4.000   Max.   :3.000   Max.   :3.000  

    Dewormed       Sterilized        Health         Quantity           Fee                State      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.000   Min.   :   0.00   Min.   :41324  
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:   0.00   1st Qu.:41326  
 Median :1.000   Median :2.000   Median :1.000   Median : 1.000   Median :   0.00   Median :41326  
 Mean   :1.559   Mean   :1.914   Mean   :1.037   Mean   : 1.576   Mean   :  21.26   Mean   :41346  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.: 1.000   3rd Qu.:   0.00   3rd Qu.:41401  
 Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :20.000   Max.   :3000.00   Max.   :41415  

                              RescuerID        VideoAmt                                                  Description
fa90fa5b1ee11c86938398b60abc32cb:  459   Min.   :0.00000   For Adoption                                        :  164
aa66486163b6cbc25ea62a34b11c9b91:  315   1st Qu.:0.00000   Dog 4 Adoption                                      :   54
c00756f2bdd8fa88fc9f07a8309f7d5d:  231   Median :0.00000   Cat for adoption                                    :   25
b53c34474d9e24574bcec6a3d3306a0d:  228   Mean   :0.05676   Friendly                                            :   20
ee2747ce26468ec44c7194e7d1d9dad9:  156   3rd Qu.:0.00000   Dog for adoption                                    :   18
95481e953f8aed9ec3d16fc4509537e8:  134   Max.   :8.00000   Please feel free to contact us : Stuart:   18
(Other)                           :13470                     (Other)                                             :14694
      PetID          PhotoAmt      AdoptionSpeed  
0008c5398:    1   Min.   : 0.000   Min.   :0.000  
000a290e4:    1   1st Qu.: 2.000   1st Qu.:2.000  
000fb9572:    1   Median : 3.000   Median :2.000  
0011d7c25:    1   Mean   : 3.889   Mean   :2.516  
00156db4a:    1   3rd Qu.: 5.000   3rd Qu.:4.000  
001a1aaad:    1   Max.   :30.000   Max.   :4.000  
(Other)  :14987                                   
> 
```

Figure 6: Describe statement in R for the variable statistics.

```
> describe(mydata)
              vars     n     mean       sd median  trimmed     mad   min   max range  skew kurtosis    se
Type             1 14993     1.46     0.50      1     1.45    0.00     1     2     1  0.17    -1.97  0.00
Name*            2 14993  4213.14  2765.73   4278  4197.38 3639.78     1  9061  9060  0.01    -1.24 22.59
Age              3 14993    10.45    18.16      3     5.98    2.97     0   255   255  3.76    20.76  0.15
Breed1           4 14993   265.27    60.06    266   279.51   60.79     0   307   307 -2.22     4.84  0.49
Breed2           5 14993    74.01   123.01      0    54.14    0.00     0   307   307  1.14    -0.61  1.00
Gender           6 14993     1.78     0.68      2     1.72    1.48     1     3     2  0.31    -0.86  0.01
Color1           7 14993     2.23     1.75      2     1.87    1.48     1     7     6  1.47     1.01  0.01
Color2           8 14993     3.22     2.74      2     3.15    2.97     0     7     7  0.19    -1.51  0.02
Color3           9 14993     1.88     2.98      0     1.48    0.00     0     7     7  1.01    -0.90  0.02
MaturitySize    10 14993     1.86     0.55      2     1.84    0.00     1     4     3  0.01     0.46  0.00
FurLength       11 14993     1.47     0.60      1     1.39    0.00     1     3     2  0.89    -0.21  0.00
Vaccinated      12 14993     1.73     0.67      2     1.66    1.48     1     3     2  0.37    -0.80  0.01
Dewormed        13 14993     1.56     0.70      1     1.45    0.00     1     3     2  0.85    -0.53  0.01
Sterilized      14 14993     1.91     0.57      2     1.89    0.00     1     3     2 -0.01     0.04  0.00
Health          15 14993     1.04     0.20      1     1.00    0.00     1     3     2  5.77    35.69  0.00
Quantity        16 14993     1.58     1.47      1     1.20    0.00     1    20    19  4.60    34.07  0.01
Fee             17 14993    21.26    78.41      0     2.15    0.00     0  3000  3000  8.92   191.61  0.64
State           18 14993 41346.03    32.44  41326 41341.69    0.00 41324 41415    91  1.09    -0.79  0.26
RescuerID*      19 14993  2978.29  1619.09   3094  3004.17 2022.27     1  5595  5594 -0.12    -1.16 13.22
VideoAmt        20 14993     0.06     0.35      0     0.00    0.00     0     8     8  9.46   124.37  0.00
Description*    21 14993  6977.92  4027.60   6955  6969.63 5125.35     1 14033 14032  0.02    -1.20 32.89
PetID*          22 14993  7497.00  4328.25   7497  7497.00 5556.78     1 14993 14992  0.00    -1.20 35.35
PhotoAmt        23 14993     3.89     3.49      3     3.27    2.97     0    30    30  2.86    12.64  0.03
AdoptionSpeed   24 14993     2.52     1.18      2     2.55    1.48     0     4     4 -0.16    -1.14  0.01
```

Figure 7: Datatypes in Python using dtypes command.

```
## Determinng the datatypes from the file
print(r_df.dtypes)
```

```
Type              int64
Name             object
Age               int64
Breed1            int64
Breed2            int64
Gender            int64
Color1            int64
Color2            int64
Color3            int64
MaturitySize      int64
FurLength         int64
Vaccinated        int64
Dewormed          int64
Sterilized        int64
Health            int64
Quantity          int64
Fee               int64
State             int64
RescuerID        object
VideoAmt          int64
Description      object
PetID            object
PhotoAmt        float64
AdoptionSpeed     int64
dtype: object
```

Figure 8: P-Value, Logworth and RSquare values of the Input variables in JMP

Target Variable	Attribute	Count	PValue	LogWorth	FDR LogWorth	RSquare
AdoptionSpeed	Breed1	14993	5.06383E-40	39.2955212	38.06507231	0.011628
AdoptionSpeed	Age	14993	5.70687E-35	34.2436019	33.31418294	0.010102
AdoptionSpeed	Type	14993	4.34343E-29	28.3621675	27.73377862	0.008325
AdoptionSpeed	FurLength	14993	3.79565E-29	28.4207141	27.73377862	0.008342
AdoptionSpeed	Sterilized	14993	1.38603E-24	23.8582288	23.32674991	0.006963
AdoptionSpeed	Quantity	14993	1.23178E-14	13.9094661	13.45716839	0.00396
AdoptionSpeed	Vaccinated	14993	4.64665E-13	12.3328602	11.94750927	0.003486
AdoptionSpeed	Gender	14993	1.65539E-12	11.7810998	11.45374082	0.00332
AdoptionSpeed	MaturitySize	14993	2.23928E-08	7.64989214	7.373685729	0.002084
AdoptionSpeed	Color1	14993	6.18623E-08	7.20857409	6.978125167	0.001953
AdoptionSpeed	Color2	14993	2.29844E-06	5.63856603	5.449509797	0.001488
AdoptionSpeed	Health	14993	0.000320493	3.49418187	3.342914195	0.000863
AdoptionSpeed	Breed2	14993	0.022448763	1.64880758	1.532302015	0.000348
AdoptionSpeed	Dewormed	14993	0.10579015	0.97555477	0.911432288	0.000174
AdoptionSpeed	State	14993	0.108195712	0.96578995	0.911432288	0.000172
AdoptionSpeed	Color3	14993	0.390459723	0.40842376	0.382094819	4.92E-05
AdoptionSpeed	Fee	14993	0.62286522	0.20560592	0.205605919	1.61E-05