DSC 530 Final

Chase Lemons

11/14/2018

Agenda

- Statistical Question
- Data Set Used
- 5 Variables
    - Histograms
    - Descriptive Statistics
- PMF
- CDF
- Analytical Distribution
- Scatter Plots
- Hypothesis Testing
- Regression Analysis

Statistical Question?

- Do black cats have the same outcomes as non black cats.

Data Set Used

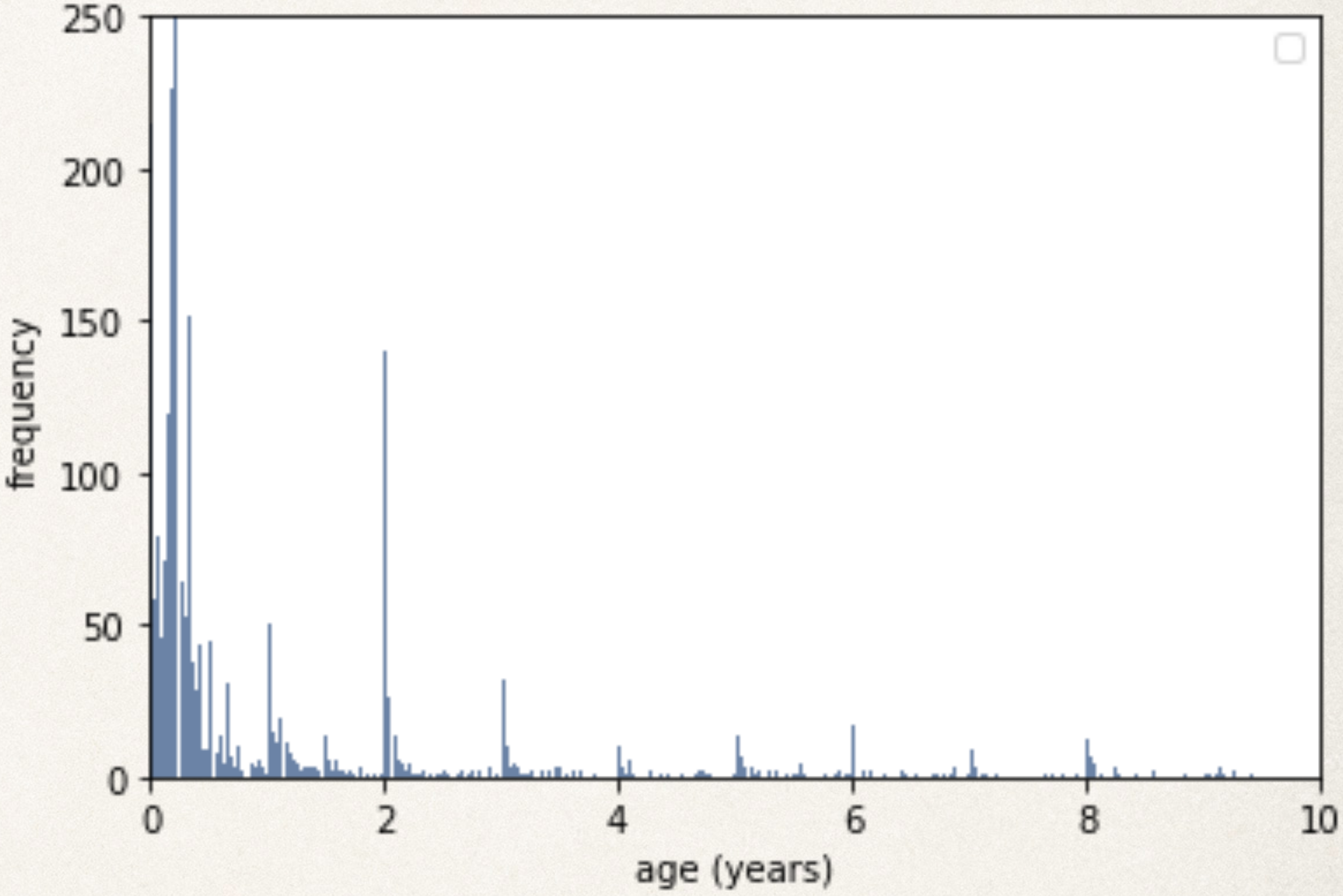
- Austin Animal Center Shelter Outcomes
  - This data set came from Kaggle. It has 37 variables and around 30,000 records

5 Variables

- Cat Age
  - Calculated by taking the difference of date of birth and the date time the animal came to the shelter. This field is numerical.
- Breeds
  - This is a categorical field.
- Color
  - Created another field called color flag from this where 0 = not black and 1 = black for the cats. This is a categorical field.
- Outcome
  - Created another field called outcome type num where the outcomes are labeled 0 - 8. This is a categorical field.
- Sex
  - Created another field called sex_flag where the outcomes are 0 and 1. This is a categorical field.

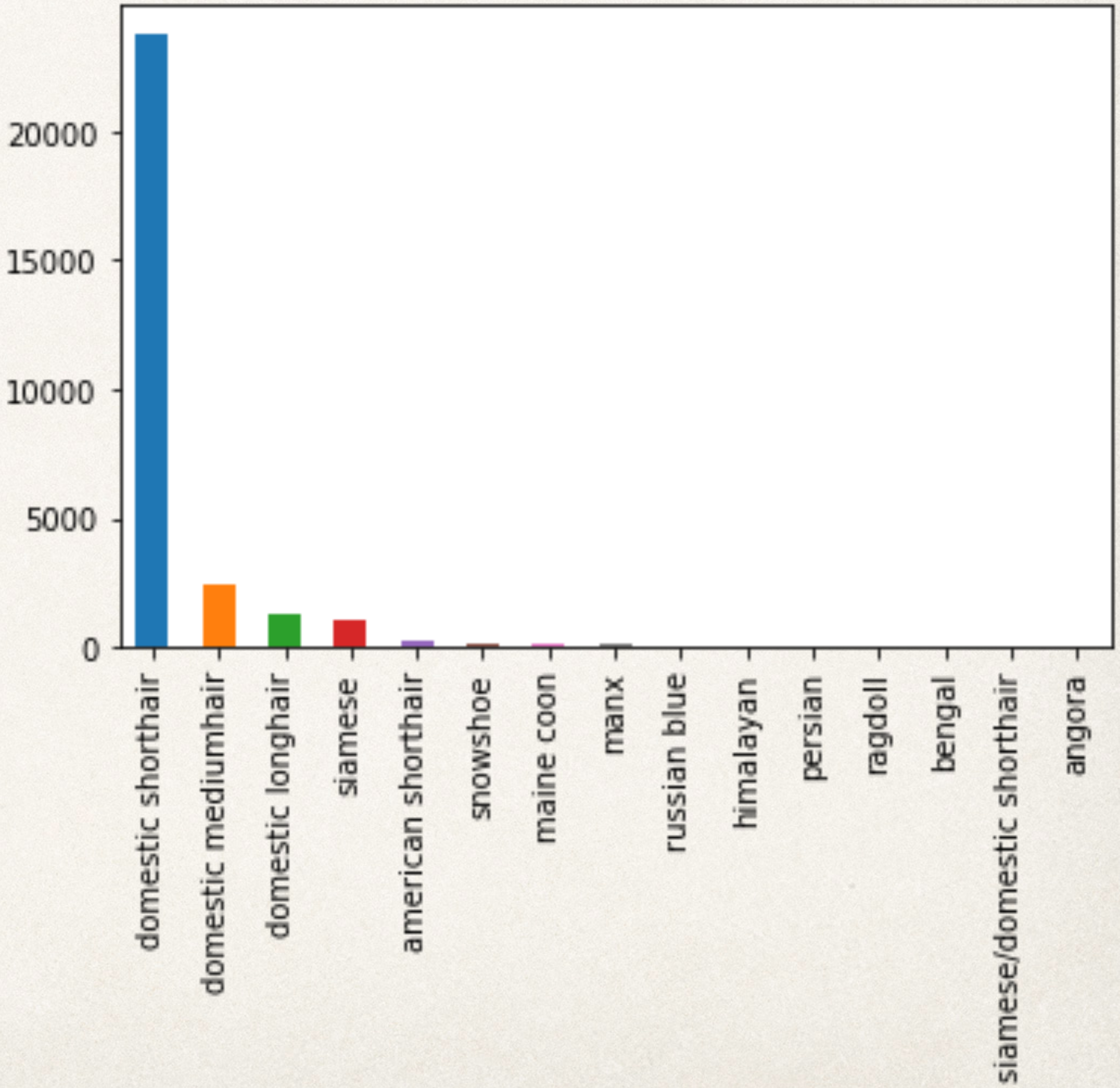
Cat Age

- Cat age histogram is to the right. This shows that there is a high amount of young cats and less older cats. There are spikes around whole and half numbers, which could potentially be explained by cats being classified by their age by estimating at the shelters. While the average is very low, the variance is high meaning that ages are spread quite a bit.
- Mean : 1.46
- Variance : 7.26
- Standard Deviation : 2.69



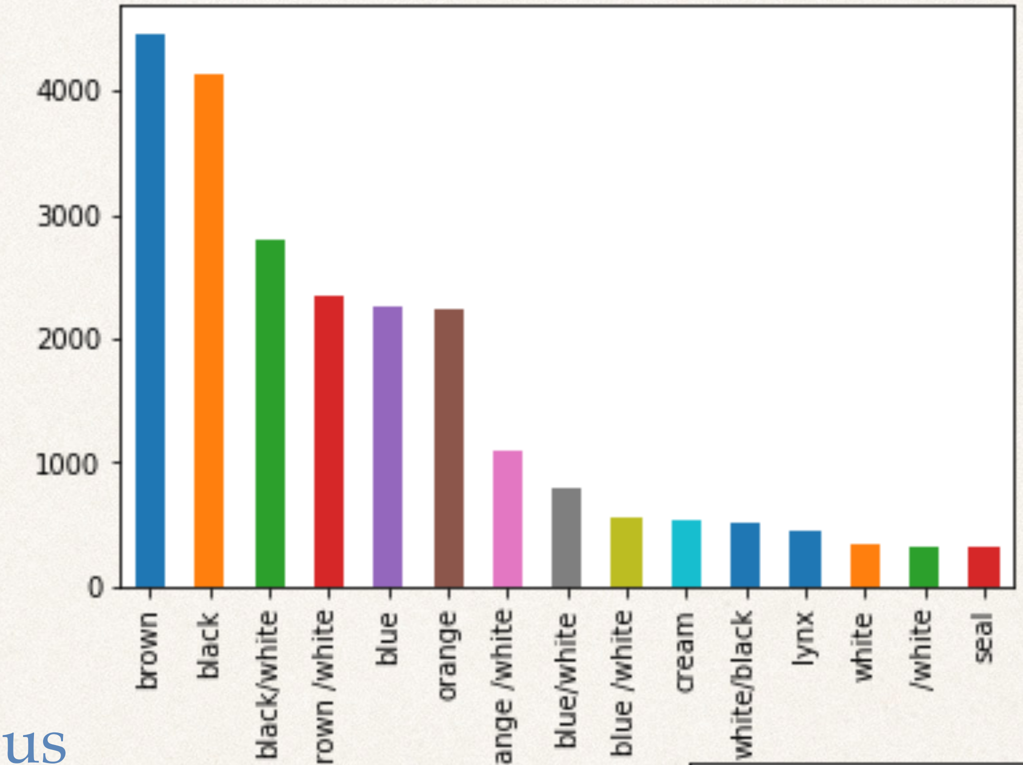
Breeds

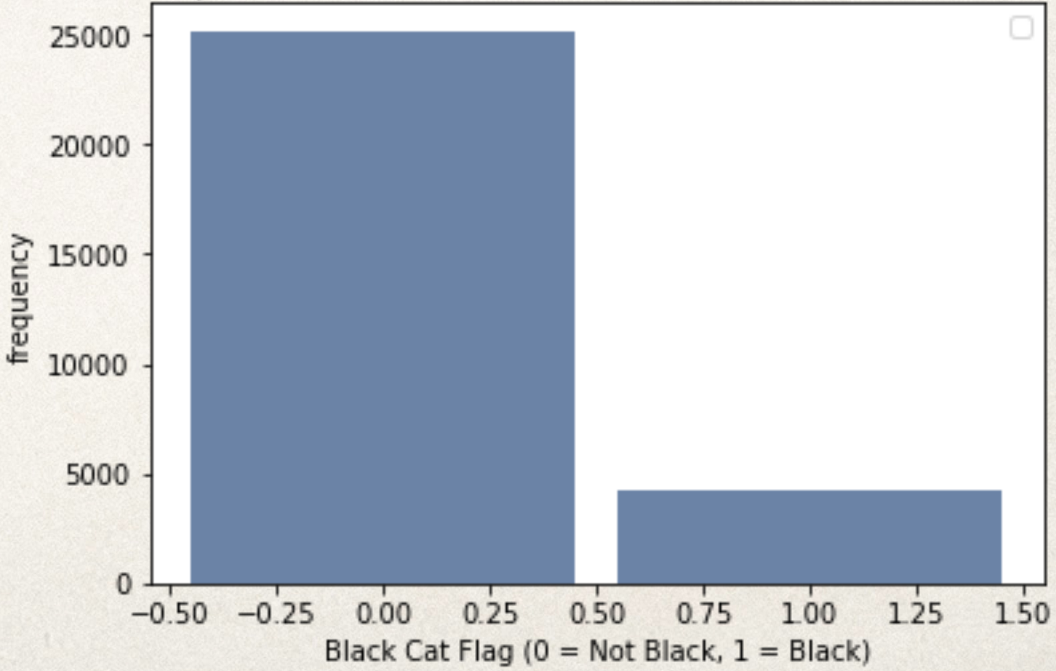
- There are many different classifications and so to the right are the top 15. Domestic shorthair is the most common breed in shelters.



Color

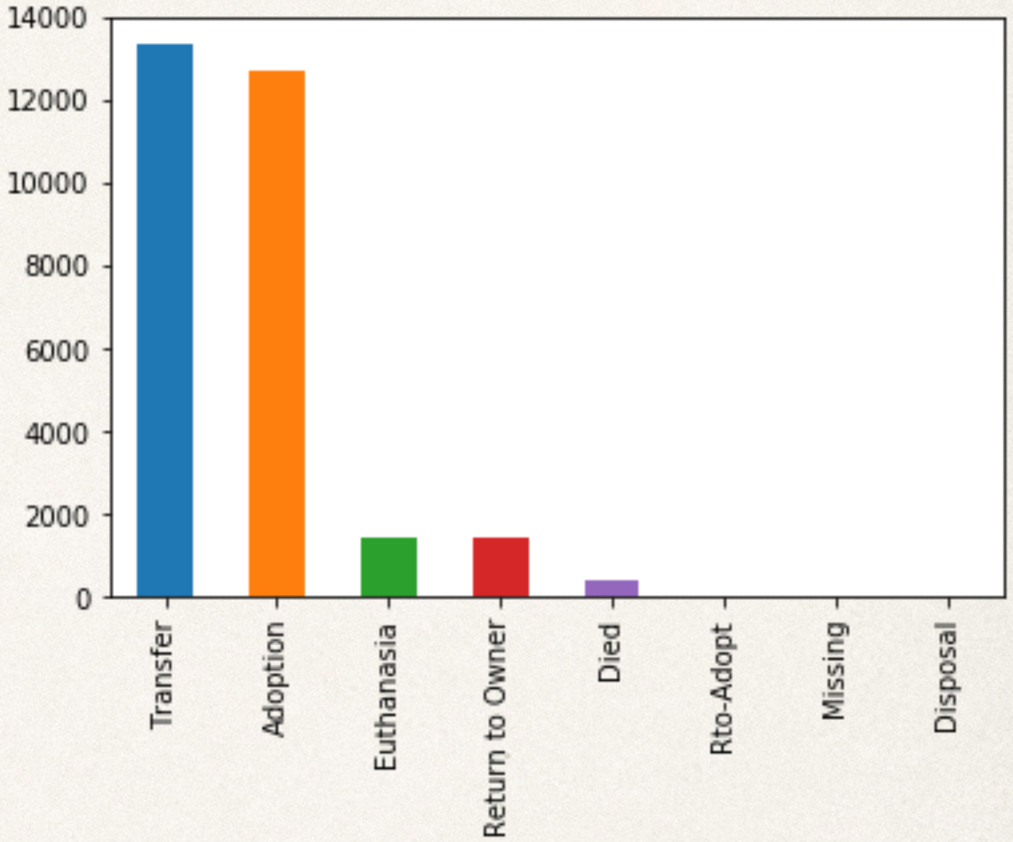
- There are many different classifications of color which can be seen in the top chart. However, the part we are interested in is black cats versus the other colored cats. Thus, we can see that there are about 25,000 other cats and about 5,000 black cats.





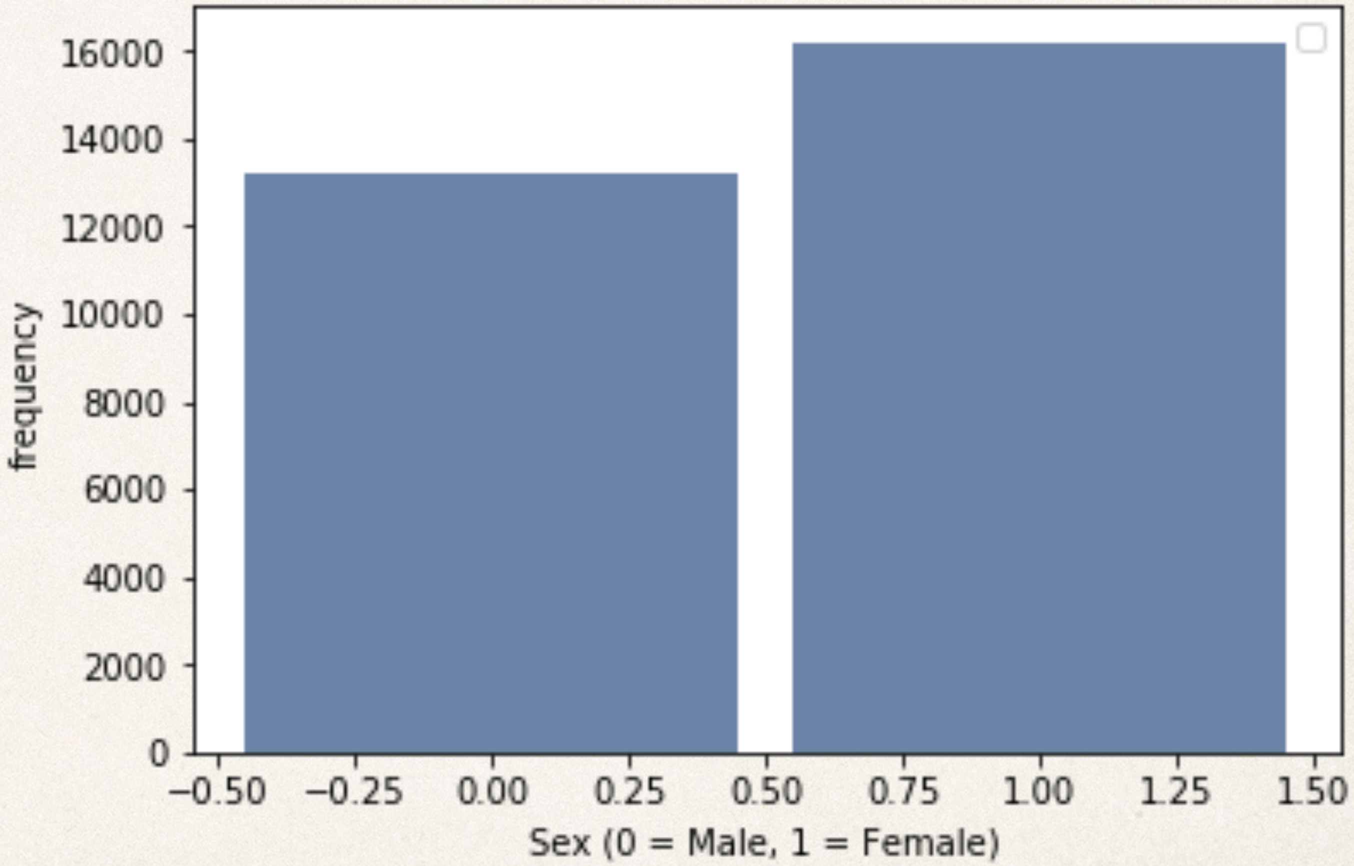
Outcome

- There are 8 outcome categories the most popular of the categories is Transfer and Adoption.
- For Reference Later on, the below categories correspond to the numbers for outcome types.
- Transfer = 0
- Adoption = 1
- Return to Owner = 2
- Died = 3
- Euthanasia = 4
- Missing = 5
- Disposal = 6
- Rto-Adopt = 7
- Everything else = 8



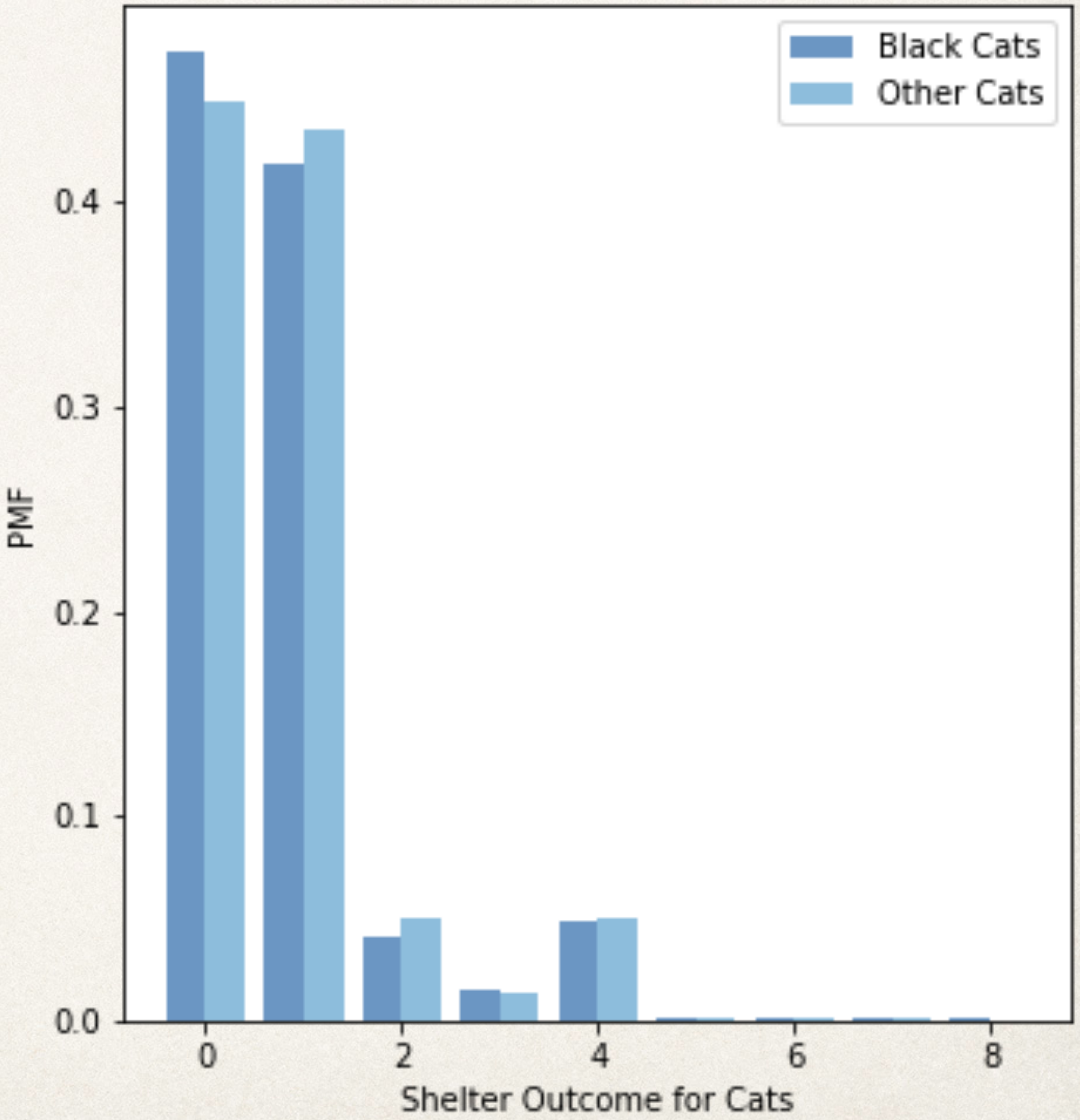
Sex

- For sex, there are two genders represented. There are more female cats in the shelters by about 3,000.



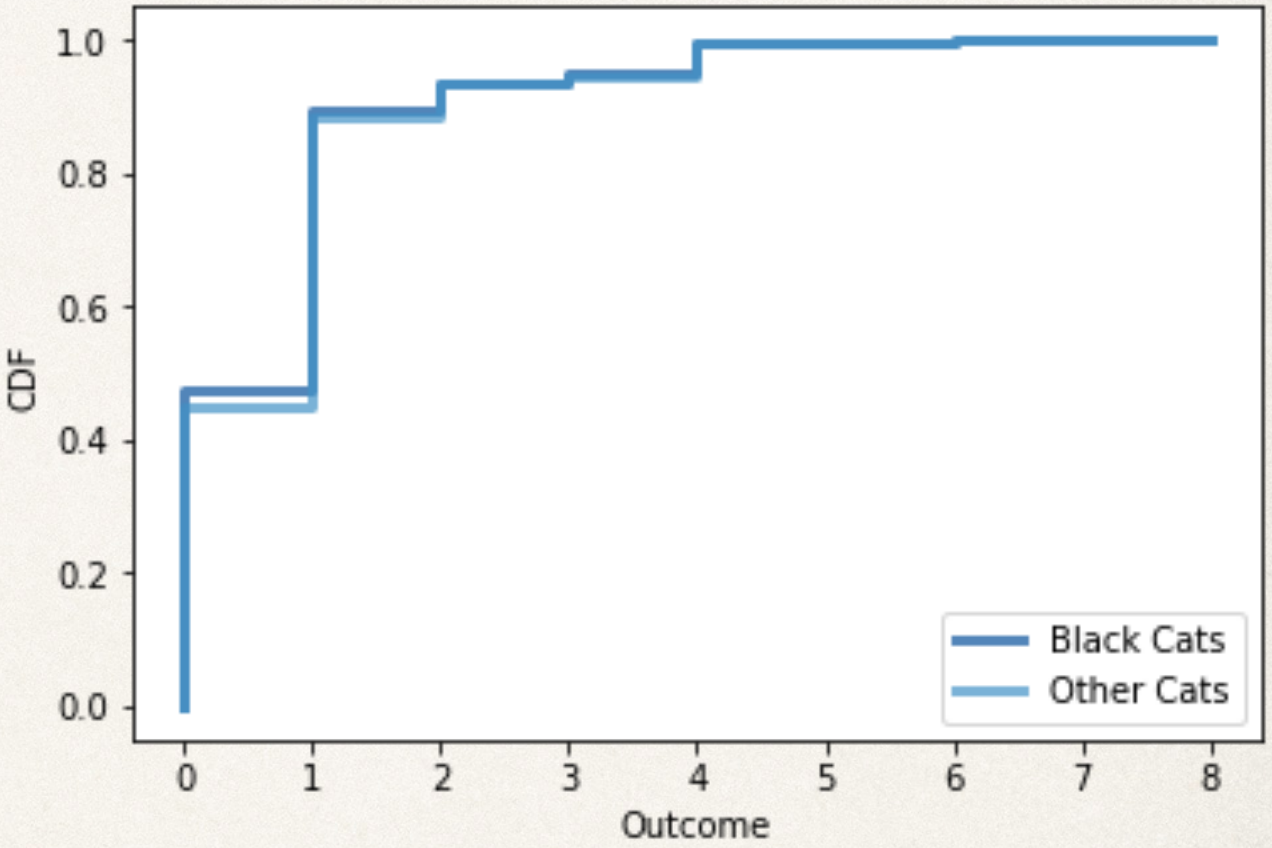
PMF

- The pmf to the right represents the comparison of shelter outcomes and a comparison of probabilities between black cats and other cats. We can see that other cats have a high probability of adoption and being returned to an owner, and euthanasia. While Black cats have a high probability oof being transferred and dying. This makes me think that Black cats may be very close to being treated the same as other cats.
- Transfer = 0
- Adoption = 1
- Return to Owner = 2
- Died = 3
- Euthanasia = 4
- Missing = 5
- Disposal = 6
- Rto-Adopt = 7
- Everything else = 8



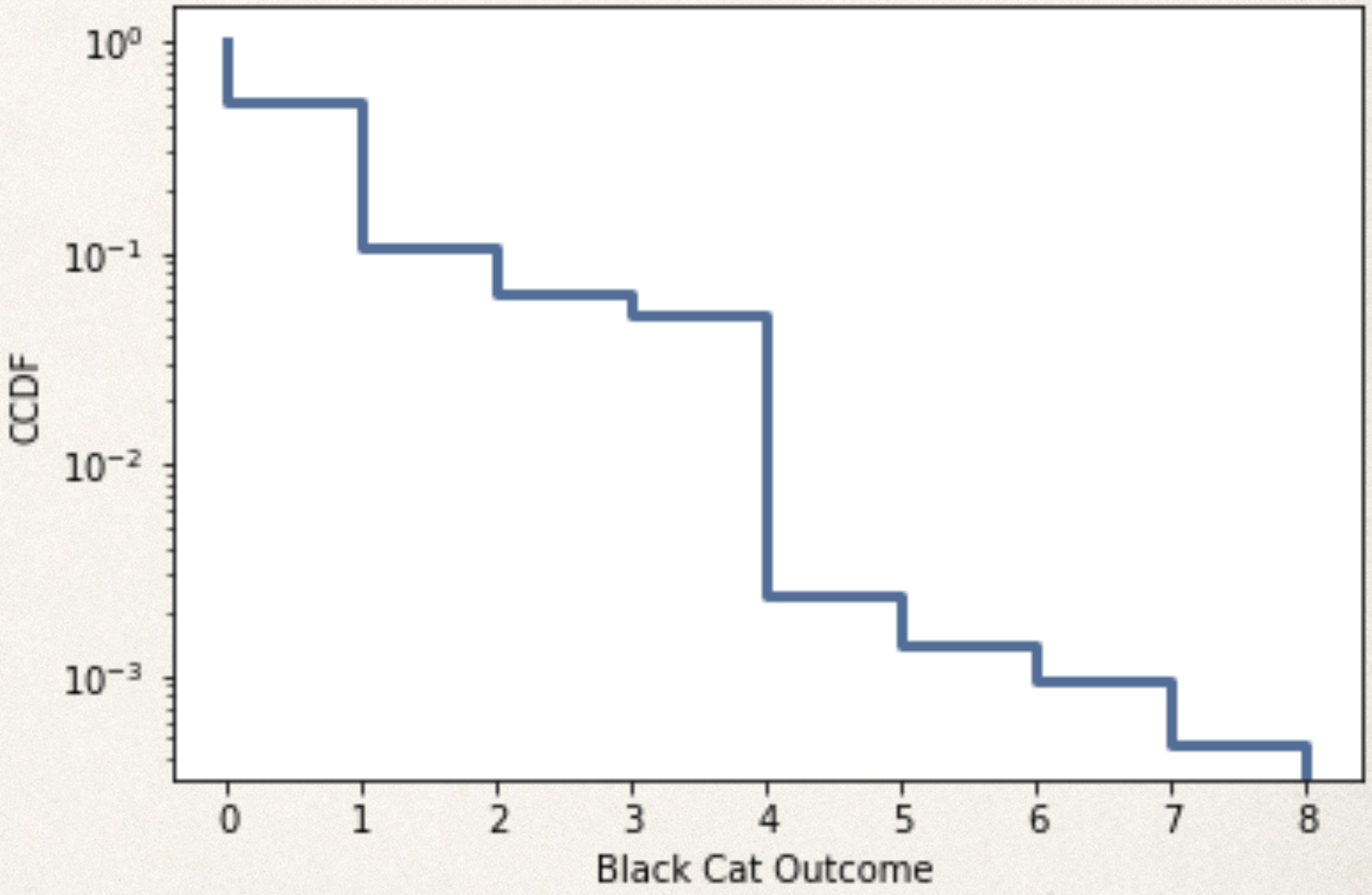
CDF

- The cdf to the right shows that the most common outcome types are transfers and adoptions. As well, we compared the cdf of black cats versus other cats and you can see that they have very similar outcomes. Making me think that black cats are treated close to equal to not black cats.
- Transfer = 0
- Adoption = 1
- Return to Owner = 2
- Died = 3
- Euthanasia = 4
- Missing = 5
- Disposal = 6
- Rto-Adopt = 7
- Everything else = 8



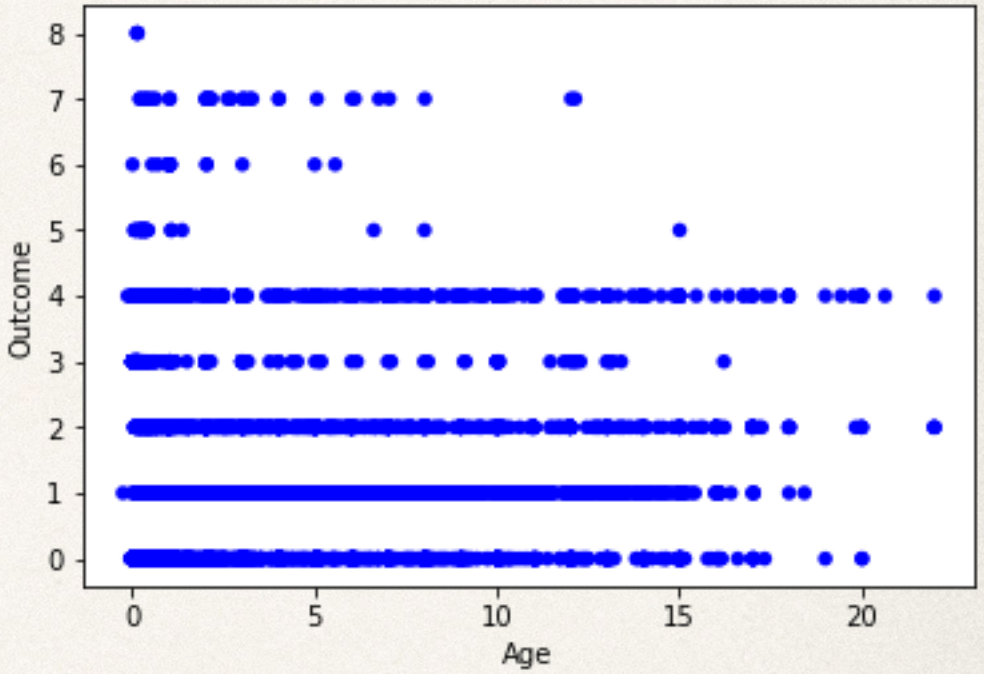
Analytical Distribution

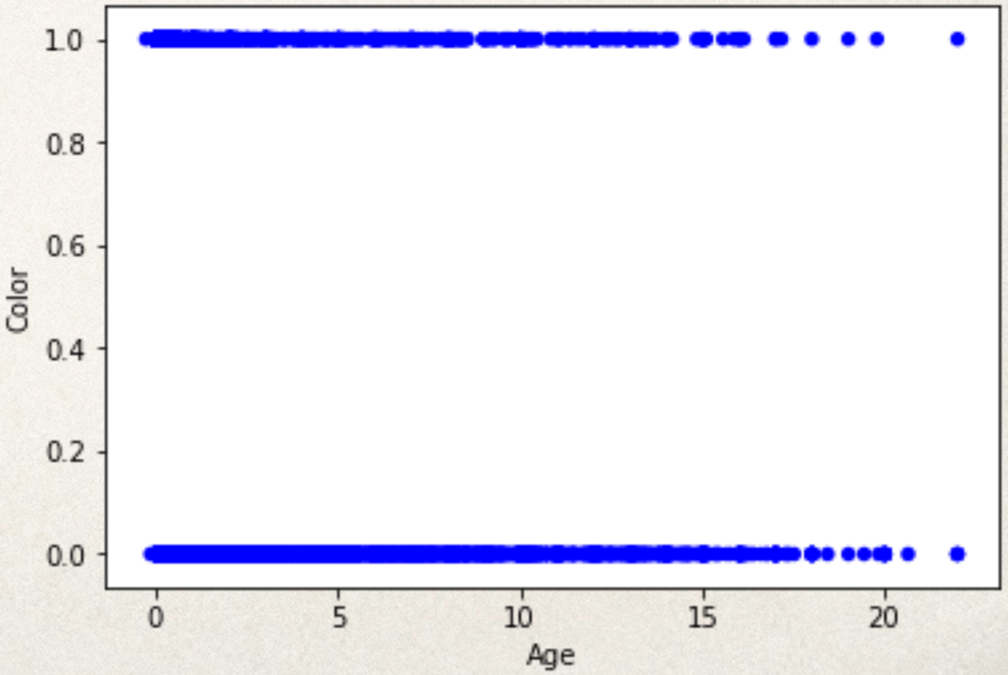
- The complimentary CDF is to the right. Data that fits an exponential distribution would have provided a straight line and thus the exponential model isn't a good model for this data. This points out that a black cat does not have equal probability for each of these outcomes.
- Transfer = 0
- Adoption = 1
- Return to Owner = 2
- Died = 3
- Euthanasia = 4
- Missing = 5
- Disposal = 6
- Rto-Adopt = 7
- Everything else = 8



Scatter Plots

- Since the data is categorical except for age, scatter plots aren't necessarily the best choice for data visualization. However to the right we have Age versus Outcome and Age versus Color. Outcomes 1,2, and 4 are highly populated and as well we already saw with the data that there are a lot of young cats hence why the younger ages are more densely populated. With the Color versus Age, there is a pretty even spread of age and color.
- Transfer = 0
- Adoption = 1
- Return to Owner = 2
- Died = 3
- Euthanasia = 4
- Missing = 5
- Disposal = 6
- Rto-Adopt = 7
- Everything else = 8





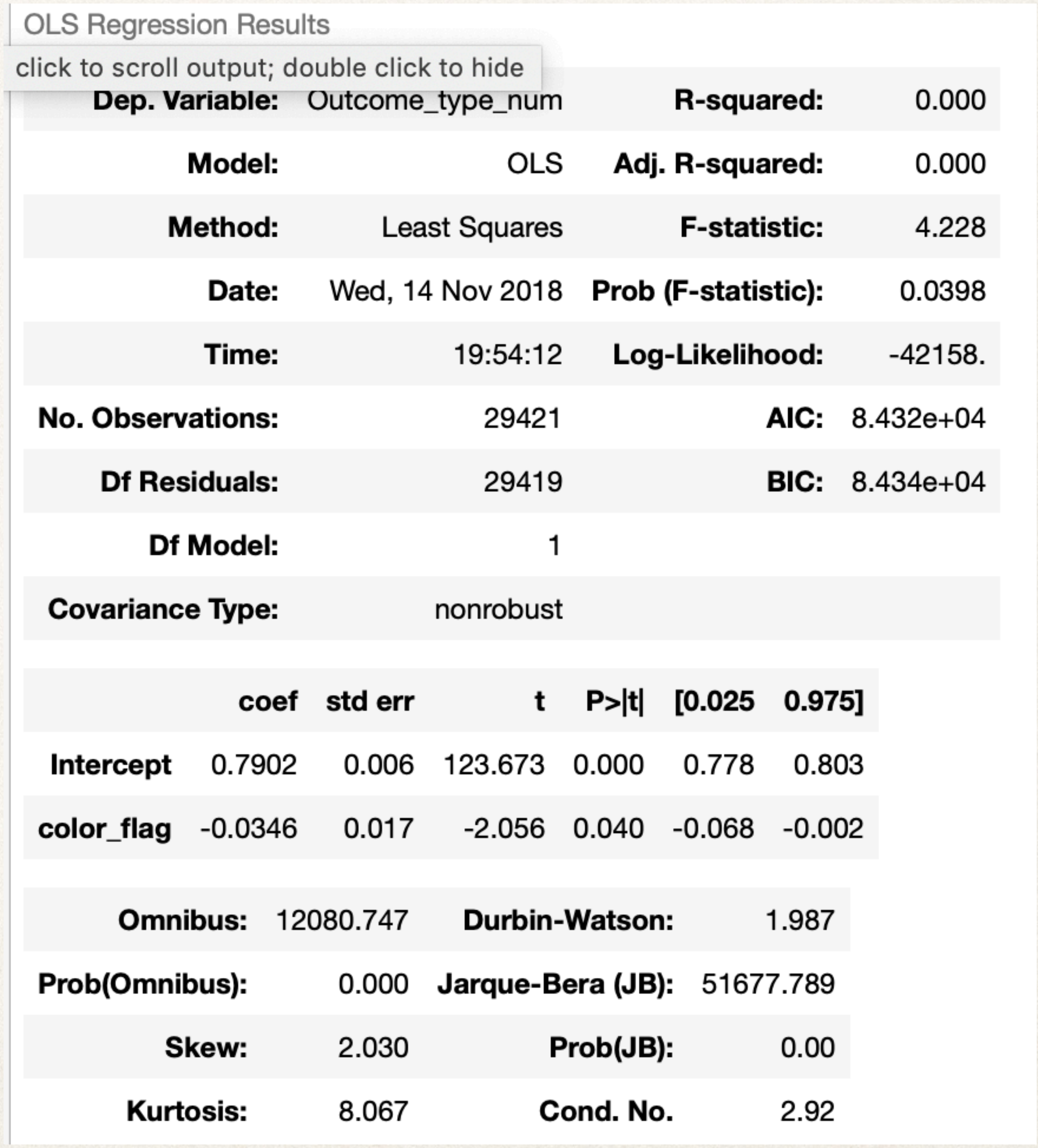
Hypothesis Testing

- For the hypothesis test, I decided to test the difference in means of outcome types. My thought process was that if black cats and non black cats have the same probability of outcomes and hence treated the same, then if you were to take an average of the numeric outcomes, then their averages should be equal. So for my null hypothesis, I assumed that the average of the outcomes for black cats and not black cats would be the same.The p-value is less than 0.05 and thus statistically significant and we reject the null hypothesis. The conclusion we draw is that, black cats have different outcomes than not black cats.

P Value = 0.041

Regression

- Doing a regression analysis, we know this isn't a good model. To start, Our R-Squared is 0. This means our model explains 0% of the variation in outcome. Our t values based on their size don't indicate a good model as well.



```
OLS Regression Results
click to scroll output; double click to hide
   Dep. Variable:   Outcome_type_num          R-squared:      0.000
           Model:                OLS     Adj. R-squared:      0.000
          Method:      Least Squares        F-statistic:      4.228
            Date:   Wed, 14 Nov 2018  Prob (F-statistic):    0.0398
            Time:           19:54:12     Log-Likelihood:    -42158.
No. Observations:              29421                AIC:  8.432e+04
    Df Residuals:              29419                BIC:  8.434e+04
        Df Model:                  1
 Covariance Type:          nonrobust

               coef  std err        t   P>|t|  [0.025   0.975]
 Intercept   0.7902    0.006  123.673   0.000   0.778    0.803
color_flag  -0.0346    0.017   -2.056   0.040  -0.068   -0.002

      Omnibus:  12080.747    Durbin-Watson:      1.987
Prob(Omnibus):      0.000  Jarque-Bera (JB):  51677.789
         Skew:      2.030         Prob(JB):       0.00
     Kurtosis:      8.067         Cond. No.       2.92
```