Capstone final

Harry Secor

12/14/2022

Introduction

The data set that I am working with came from a paper from Joanna Ilska, et al. that asks how does a dogs traits effects its personality. The data is from a 102 question questionnaire from another paper by Sarah E.Lofgren, et al. that asks questions about dog behavior and generates a number between 1-5 for each behavioral trait. The data consist of 1975 such responses by owners of Labrador retriever owners who's dogs are registered at the UK kennel club. There are 7 predictor variables coat color, gender status, age, job, housing, exercise, and health. all the predictor variables are categorical except age and they range between 1-2 to 1-4. There are also two sets of response variables, each response being a number from 1-5,, with set A being Agitated (level of agitation when ignored), Attenseek (attention), Bark (barking tendency), Excite (excitement), Fetch (how well does the dog fetch), HO_Fear (Human and Object Fear), NoiseF (Noise Fear), NO_Agg (Non-owner-directed Aggression), O_Agg (Owner-directed Aggression), SepAnx (Separation Anxiety), Train (trainability), UnBeh (Unusual Behavior). Set B I did not find the meaning of and also did not use. The reason I chose this data set is because I though it would be fun to work with dogs and nature vs nurture is always fascinating

Analysis

the data is from a study in whether a dogs genetics has an impact on its behavior or if it is more based on the actions of the dogs owners. For the tested behavioral traits the more important factor is how the dogs were nurtured.

find male vs female by Gender_status %% 2 need to do this for later before Gender_status becomes a factored value

```
DOG$Fixed <- DOG$Gender_status %% 2
```

I used summary to look at the data and see the ranges of each of the categories but mainly paid attention to the response categories between Agitated and UnBeh

```
summary(DOG)
```

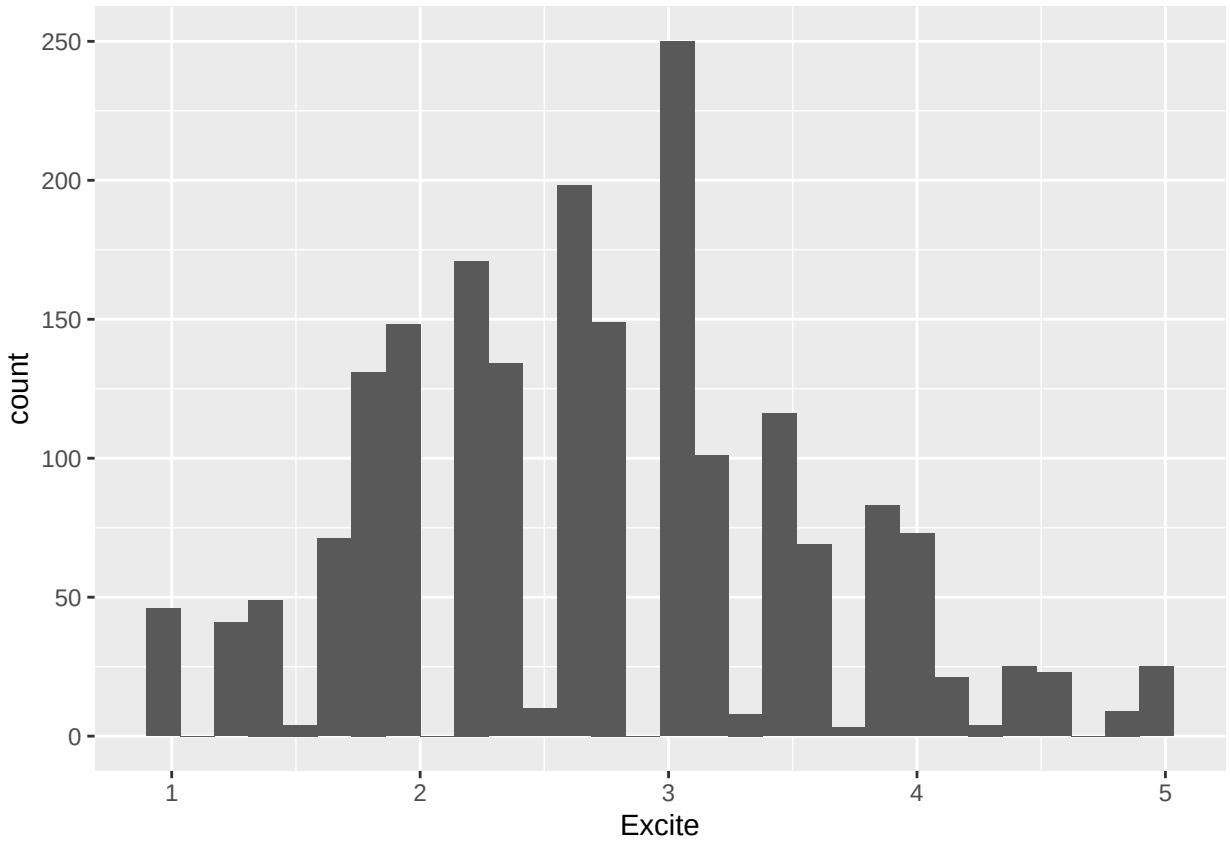
```
##   Coat_colour     Gender_status        Age            WM2
##  Min.   :0.0000   Min.   :0.000   Min.   : 760   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1680   1st Qu.:0.0000
##  Median :0.0000   Median :2.000   Median :2052   Median :1.0000
##  Mean   :0.5777   Mean   :1.728   Mean   :2042   Mean   :0.6105
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:2379   3rd Qu.:1.0000
##  Max.   :2.0000   Max.   :3.000   Max.   :3380   Max.   :2.0000
##                   NA's   :11                     NA's   :178
##      InOut            TimeEx          Health          Agitated
##  Min.   :0.0000   Min.   :1.000   Min.   :0.0000   Min.   :1.000
```

```
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.0000   Median :2.000   Median :0.0000   Median :1.500
##  Mean   :0.2682   Mean   :2.247   Mean   :0.1408   Mean   :1.853
##  3rd Qu.:0.0000   3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:2.500
##  Max.   :2.0000   Max.   :4.000   Max.   :1.0000   Max.   :5.000
##  NA's   :51       NA's   :5                        NA's   :74
##    AttenSeek          Bark           Excite          Fetch
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.667   1st Qu.:1.000   1st Qu.:2.200   1st Qu.:4.000
##  Median :3.333   Median :1.000   Median :2.600   Median :5.000
##  Mean   :3.346   Mean   :1.574   Mean   :2.711   Mean   :4.514
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.200   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :33      NA's   :20      NA's   :13      NA's   :22
##     HO_Fear           NoiseF          NO_Agg          O_Agg
##  Min.   :0.7333   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0667   1st Qu.:1.000   1st Qu.:1.143   1st Qu.:1.000
##  Median :1.2000   Median :1.000   Median :1.357   Median :1.000
##  Mean   :1.3243   Mean   :1.524   Mean   :1.487   Mean   :1.025
##  3rd Qu.:1.4667   3rd Qu.:2.000   3rd Qu.:1.739   3rd Qu.:1.000
##  Max.   :3.4667   Max.   :5.000   Max.   :3.857   Max.   :2.429
##  NA's   :5        NA's   :33      NA's   :4       NA's   :8
##      SepAnx          Train           UnBeh       HS_StrngDirAgg
##  Min.   :1.000   Min.   :1.714   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:3.833   1st Qu.:1.500   1st Qu.:1.000
##  Median :1.000   Median :4.143   Median :1.800   Median :1.250
##  Mean   :1.142   Mean   :4.117   Mean   :1.827   Mean   :1.414
##  3rd Qu.:1.250   3rd Qu.:4.571   3rd Qu.:2.100   3rd Qu.:1.667
##  Max.   :3.000   Max.   :5.000   Max.   :3.550   Max.   :3.556
##  NA's   :28      NA's   :6       NA's   :7       NA's   :26
##   HS_OwnDirAgg   HS_StrngDirFear HS_NonSocFear   HS_DogDirAggression
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :1.000   Median :1.167   Median :1.500
##  Mean   :1.024   Mean   :1.234   Mean   :1.357   Mean   :1.671
##  3rd Qu.:1.000   3rd Qu.:1.250   3rd Qu.:1.500   3rd Qu.:2.000
##  Max.   :2.375   Max.   :5.000   Max.   :3.667   Max.   :5.000
##  NA's   :21      NA's   :22      NA's   :14      NA's   :44
##  HS_DogDirFear    HS_SepRelBeh     HS_Attach        HS_Train
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.625
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.333   1st Qu.:3.250
##  Median :1.250   Median :1.000   Median :2.833   Median :3.500
##  Mean   :1.497   Mean   :1.139   Mean   :2.799   Mean   :3.448
##  3rd Qu.:1.750   3rd Qu.:1.250   3rd Qu.:3.333   3rd Qu.:3.625
##  Max.   :5.000   Max.   :2.875   Max.   :5.000   Max.   :4.667
##  NA's   :30      NA's   :33      NA's   :33      NA's   :14
##     HS_Chase        HS_Excit        HS_Pain        HS_DogRiv       HS_EnLevel
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:1.500   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.00
##  Median :2.500   Median :2.667   Median :1.333   Median :1.000   Median :3.00
##  Mean   :2.459   Mean   :2.660   Mean   :1.427   Mean   :1.223   Mean   :2.73
##  3rd Qu.:3.250   3rd Qu.:3.167   3rd Qu.:1.667   3rd Qu.:1.250   3rd Qu.:3.50
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :3.750   Max.   :5.00
##  NA's   :59      NA's   :20      NA's   :490     NA's   :132     NA's   :32
```

```
##      Fixed
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2469
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :11
```

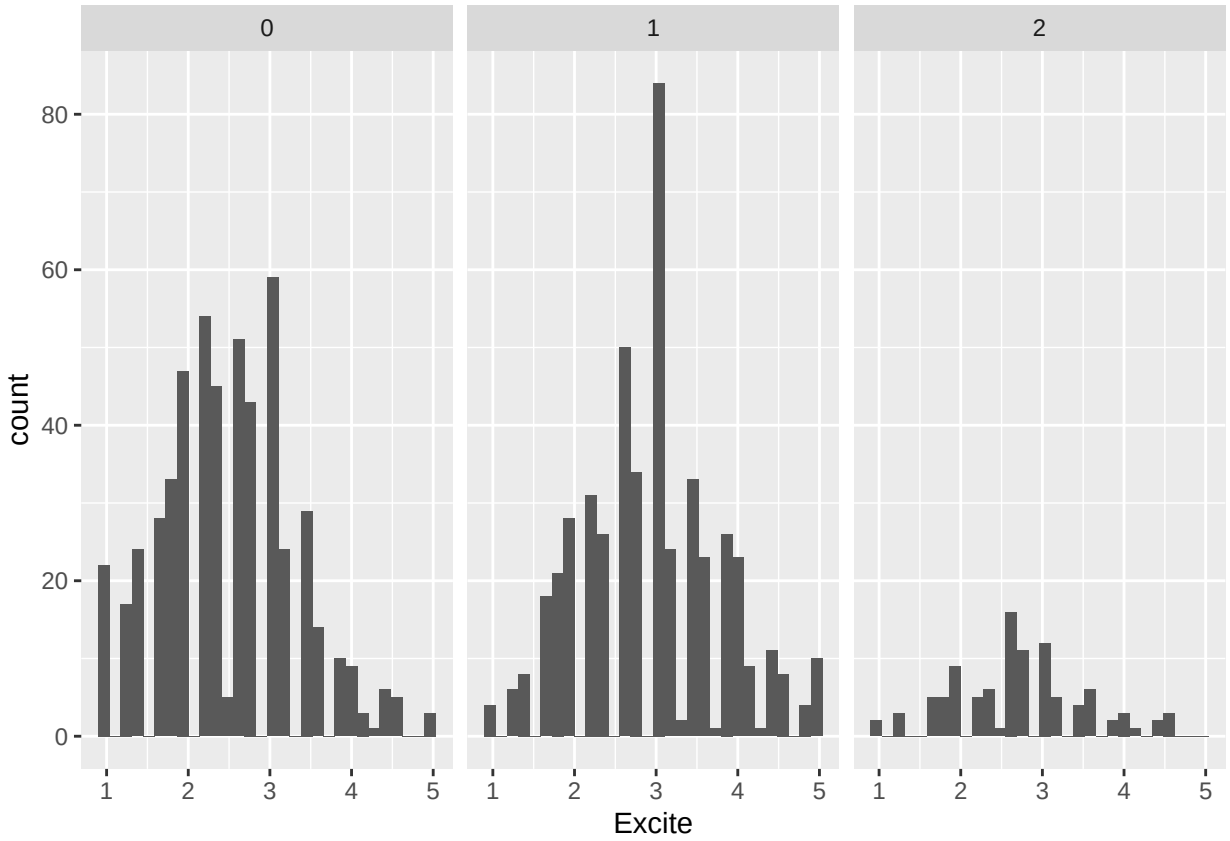
Job vs Excitability

the first two categories I want to look at are job (WM2) and excitement (Excite) to see whether training effects excitement level. WM2 is categorical with 0 = Gundog (n = 840), 1 = Pet (n = 817), and 2 = Showdog (n = 140). The paper by Joanna Ilska, et al. does not say what exactly excitement is, it is just a measurement on how excitable the dog is. Excitability is on a scale of 1 to 5 based off off six questions on a quiz by Sarah E.Lofgren et al.



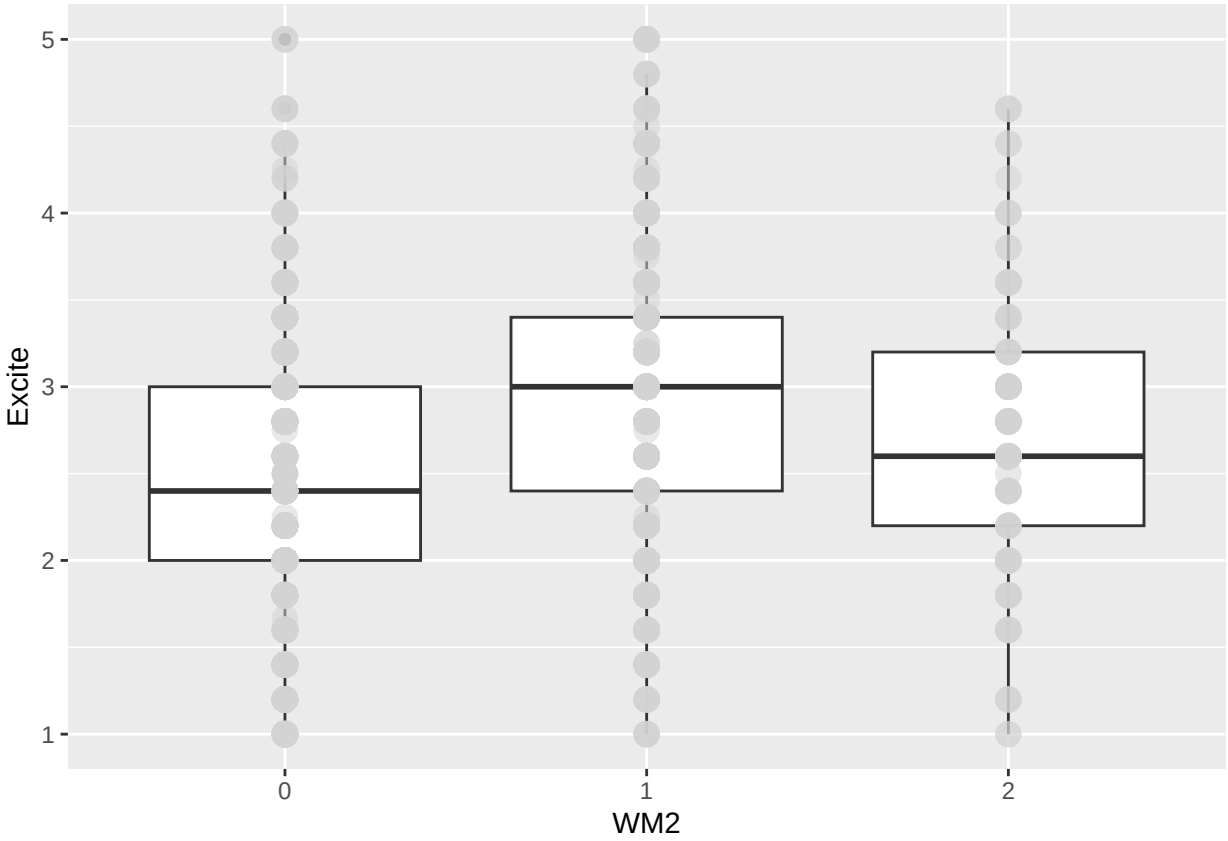
the histogram looks to be a normal distribution

histograms where x = Excite and facet_wrap job I did a little extra for this first one to look at separate histograms to look at the spread for each job.



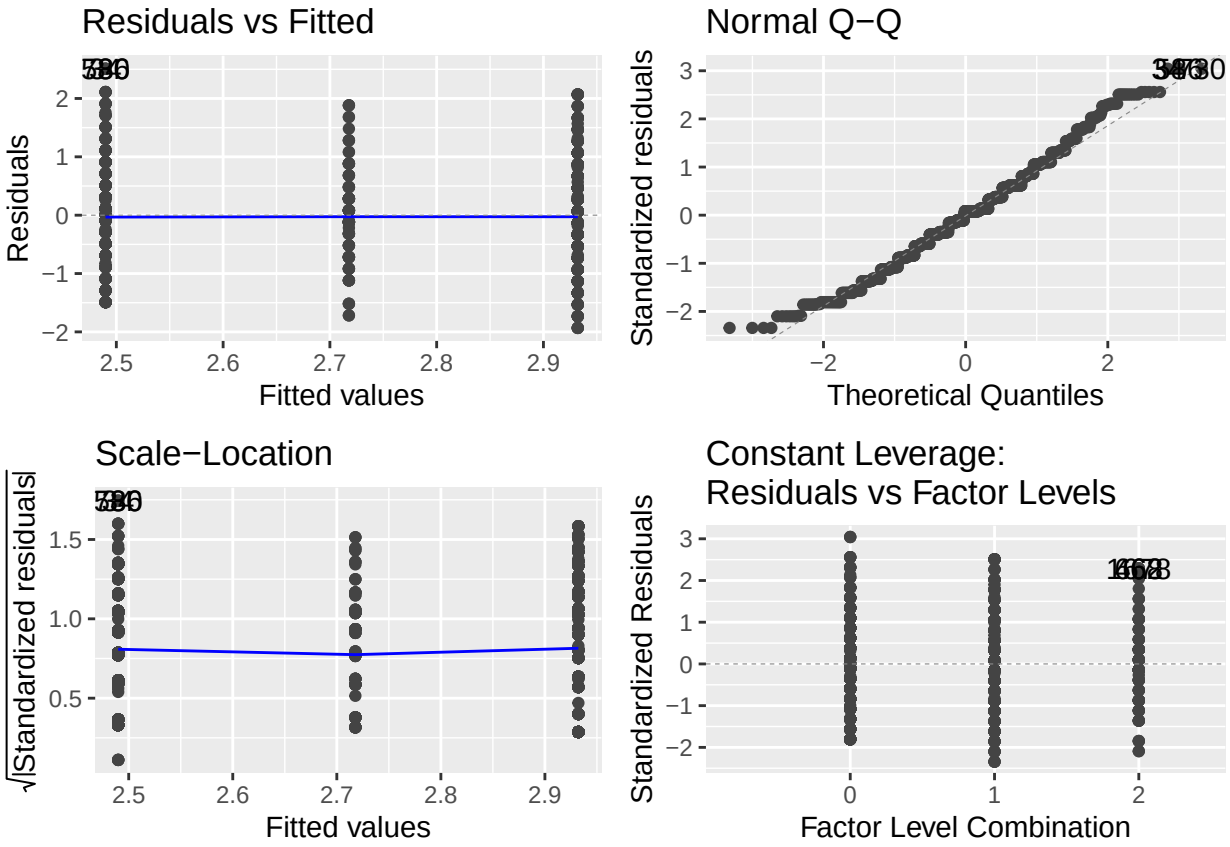
there might be a slight right skew for the gundogs and it is hard to tell for the showdogs but they look relatively good

box plot of x = WM2 and y = Excite



From the above plots pets seem to be much more excitable than either gundogs or show dogs and I think that there will be a difference between dogs that essentially live to bring happiness to their owner and dogs that have jobs. The median of pets is at about the third quartile of gundogs and show dogs are kind of in the middle of both. The job that a dog is trained for effects how excitable the dog is.

autoplot and summary of linear model with y = Excite




```
##
## Call:
## lm(formula = Excite ~ WM2, data = na.omit(DOG))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.93216 -0.53216  0.06784  0.51012  2.51012
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48988    0.03580  69.557   <2e-16 ***
## WM21         0.44228    0.05184   8.532   <2e-16 ***
## WM22         0.22794    0.08961   2.544   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8256 on 1115 degrees of freedom
## Multiple R-squared:  0.06132,    Adjusted R-squared:  0.05964
## F-statistic: 36.42 on 2 and 1115 DF,  p-value: 4.763e-16
```

The residual fitted looks very good as the blue line matches the grey dotted line and there are few outlines. Even though it looks a little funny, job is a categorical so the values should be in parallel lines like they are. The points on the normal Q-Q look nearly identical to the grey line except at the top.

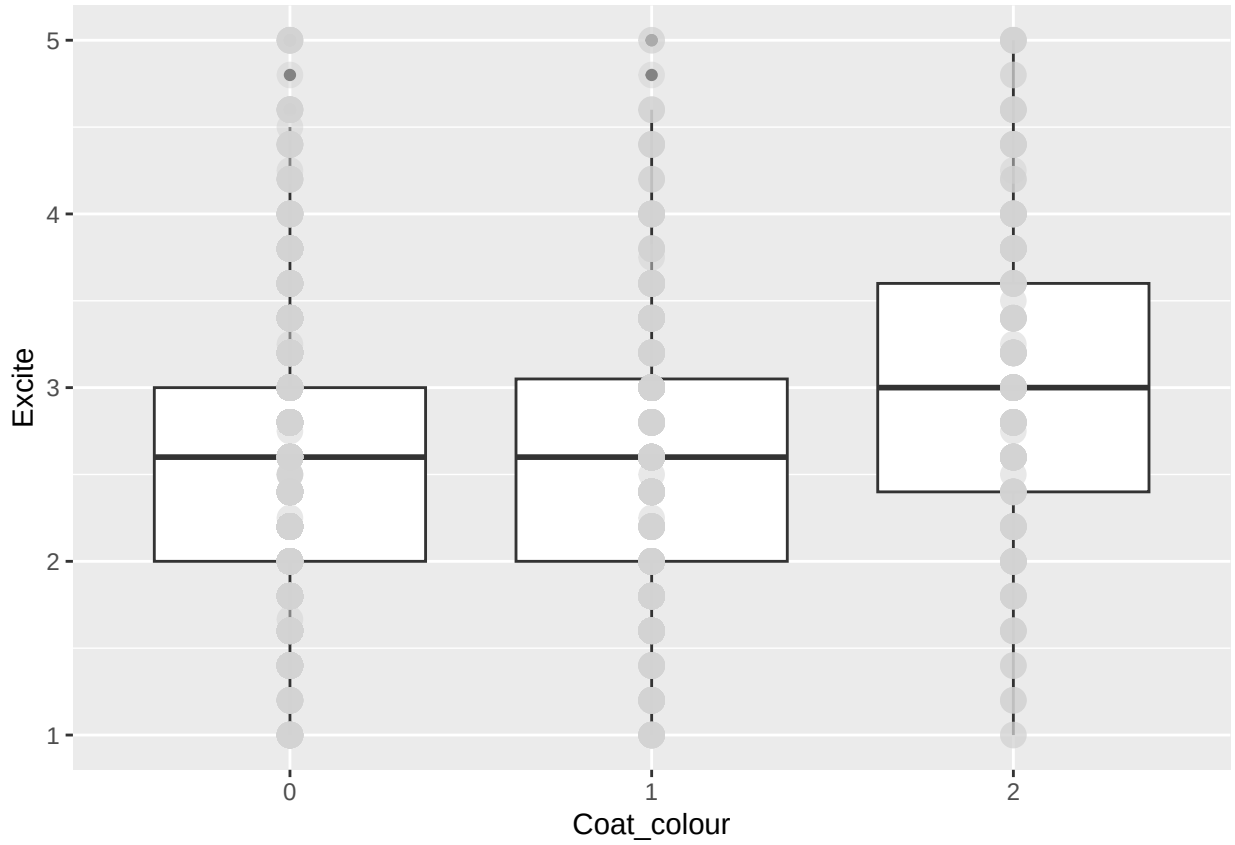
There is a significant differences between all three jobs with the biggest difference between gundog and pet and with p = 2e-16. There is a closer but still statistically significant difference between pet and show dog. We fail to reject the hypothesis that a dogs job effects its excitability. Teaching dogs different jobs does seem

have an impact on how they act.

Coat color vs Excitability

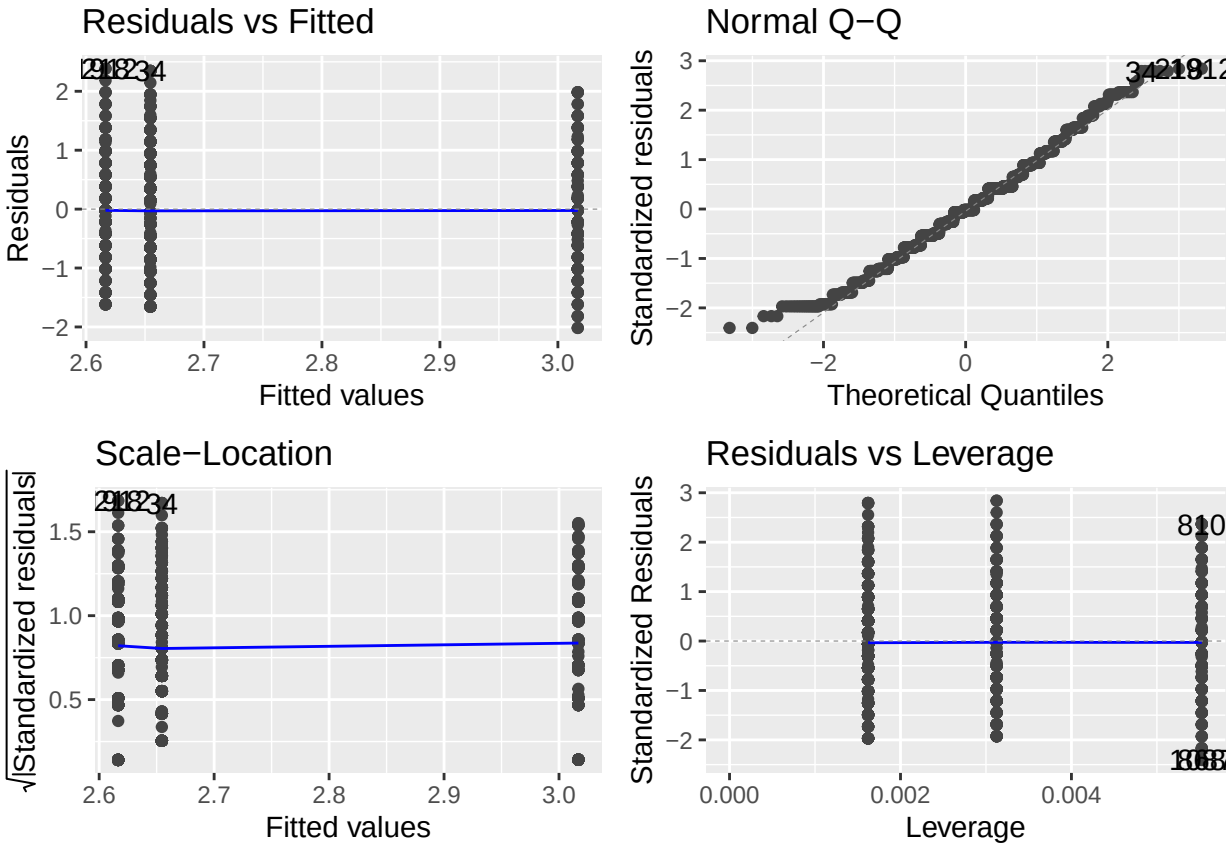
This is a very similar test to the last one because job was a more nurture variable whereas coat color is a genetic variable. There are three coat colors, 0 = black (n = 1144), 1 = yellow (n = 521), and 2 = chocolate (n = 310).

box plot of x = Coat_colour and y = Excite



We think that it is unlikely there is a significant differences between any of the coat colors. the only possibility is that chocolates are different because there mean is about the third quartile for both black and yellow labs. A dogs coat color effects its level of excitability.

autoplot and summary of linear model with x = Coat_colour and y = Excite




```
##
## Call:
## lm(formula = Excite ~ Coat_colour, data = na.omit(DOG))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01685 -0.61656 -0.01685  0.54543  2.38344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.65457    0.03385  78.431  < 2e-16 ***
## Coat_colour1 -0.03800    0.05792  -0.656    0.512
## Coat_colour2  0.36229    0.07107   5.098 4.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8407 on 1115 degrees of freedom
## Multiple R-squared:  0.02676,    Adjusted R-squared:  0.02501
## F-statistic: 15.33 on 2 and 1115 DF,  p-value: 2.712e-07
```

The residual fitted looks very good as the blue line matches the grey dotted line and there are few outlines. There might be something off with the residual fitted because two of the dot lines are very close together but that could just mean that they are similar. The points on the normal Q-Q look nearly identical to the grey line except at the top.

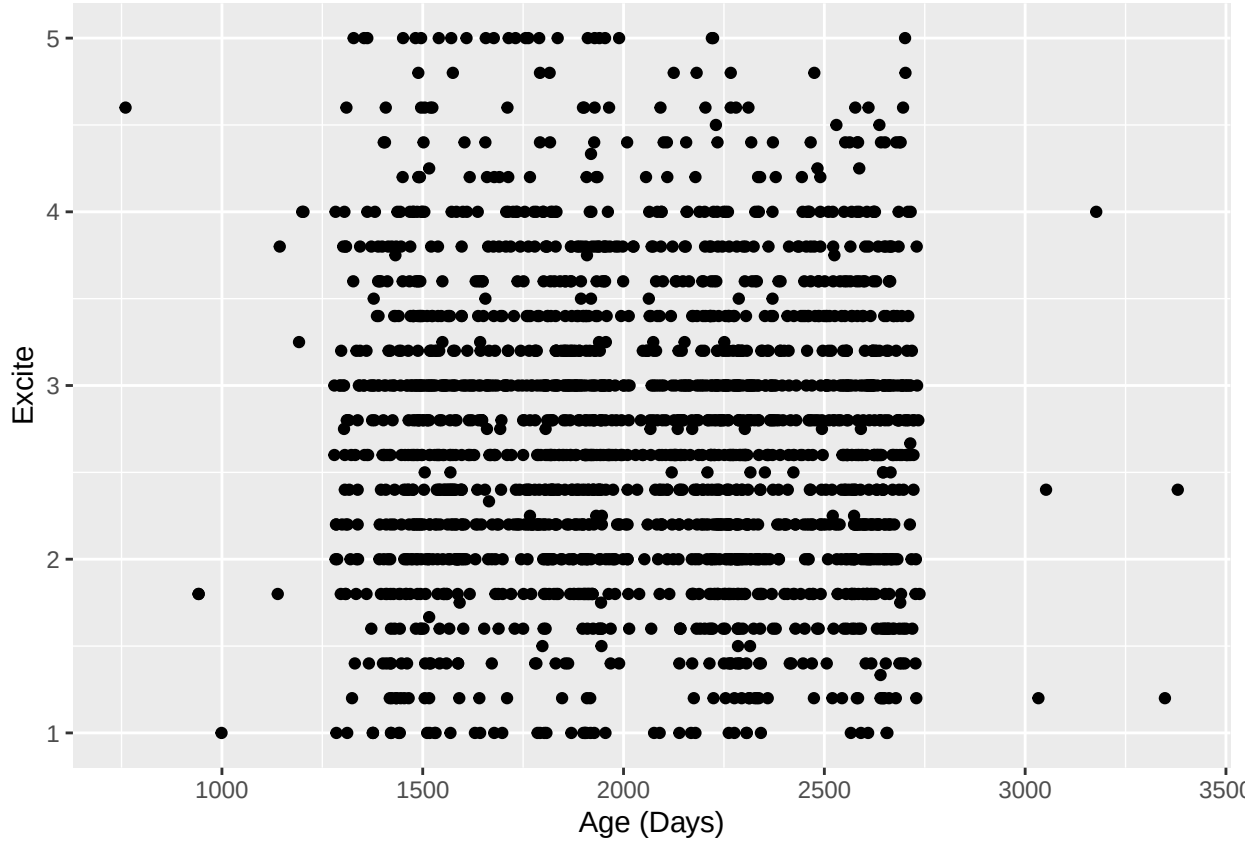
Between black labs and yellow labs there is not a significant difference in level of excitability p = 0.512 but between black labs and chocolate labs there is a significant differences in level of excitability. We have to

fail to reject the hypothesis that coat color does effect excitability in labs. We were not expecting that coat color could change how excitable dogs can be. It would likely be good to check the results of this because chocolate labs were the smallest group to see if there was another variable influencing the data.

Age vs Excite

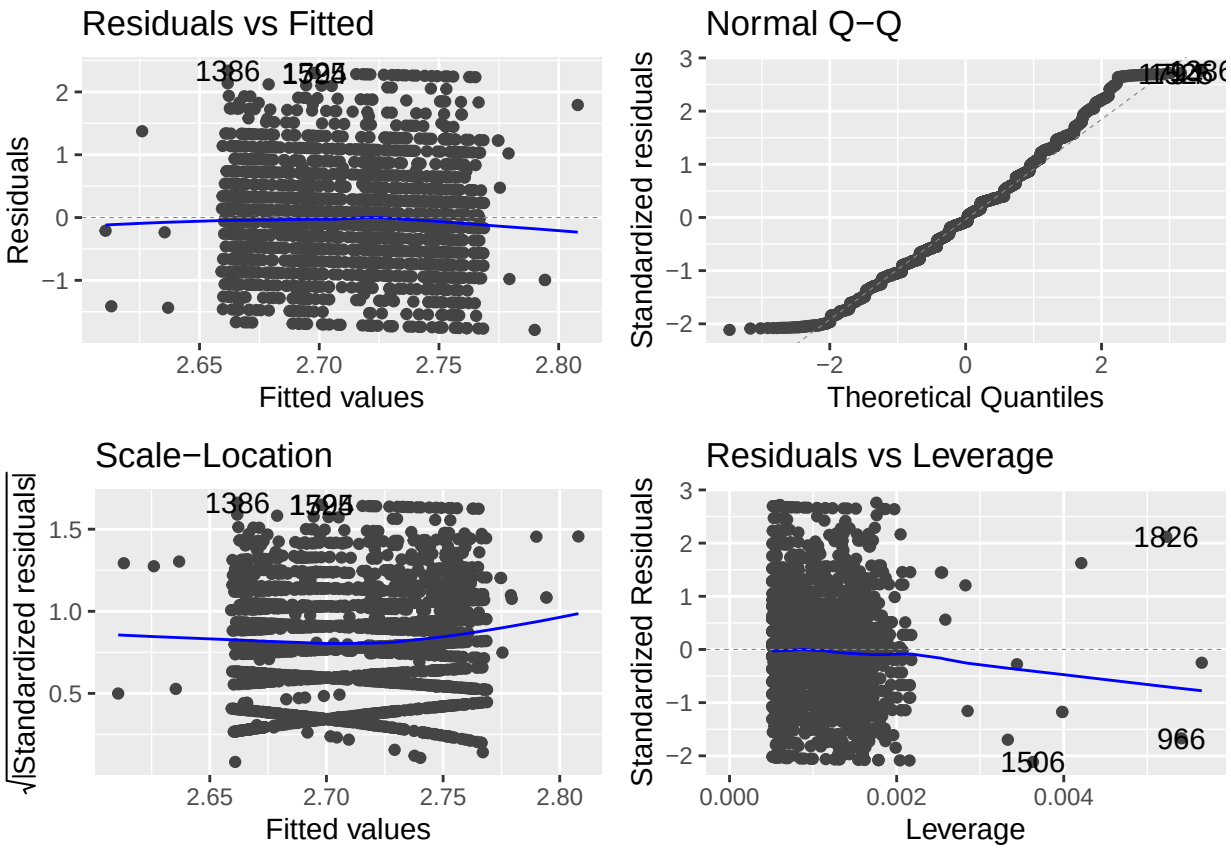
compare age to the excitement level to see what other factors could effect Excitability as a dog ages it likely gets less excited.

dot plot of x = Age and y = Excite



There is no clear pattern in the dots though after 2000 days there are only two dogs are 5 so that could be something. Seemingly most dogs are at 3 for excitability but it is hard to tell with the density off dots. Dogs get less excitable as they age.

autoplot of linear model with x = Age and y = Excite



Normal Q-Q looks very good as it is nearly perfectly on the dotted line. the Residual vs filters close to the line as well but mostly falls below. the residual vs leverage plot also veering off below the grey line after about 0.002 and most dots are behind that point so there maybe some issues and outliers. There is no good way to tell if this data is not correct so I am just going to use it.

Anova test and summary of linear model with x = Age and y = Excite

```
## Analysis of Variance Table
##
## Response: Excite
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## Age          1    1.97 1.97399  2.7458 0.09767 .
## Residuals 1960 1409.08 0.71892
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = Excite ~ Age, data = DOG)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78999 -0.56851 -0.06765  0.50688  2.33815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.865e+00  9.472e-02  30.250   <2e-16 ***
## Age         -7.529e-05  4.543e-05  -1.657   0.0977 .
```

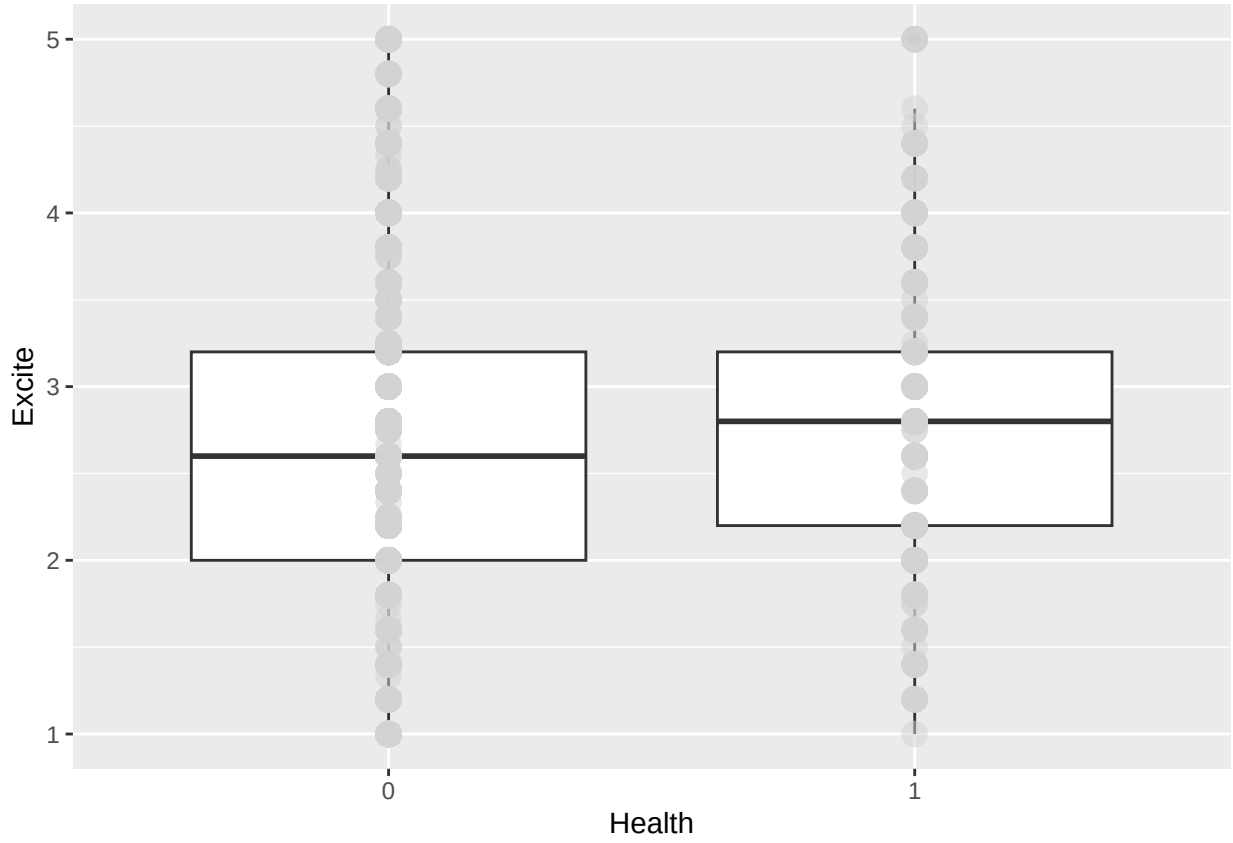
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8479 on 1960 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.001399,   Adjusted R-squared:  0.0008895
## F-statistic: 2.746 on 1 and 1960 DF,  p-value: 0.09767
```

There is no significant difference between the age of a do and its level of excitability (p = 0.0977). We reject the hypothesis that age effects level of excitability in dogs. There are some issues with this result as the autoplot had issues with several of the plots and the R^2 is very small (R-squared = 0.001399).

Health vs Excitability

Is there was another variable that could effect how excitable a dog is like health of a do to see if illness or injury made the dog less excitable. 0 = some health issues (n = 1697) and 1 = no heath issues (n = 278)

box plot of x = Health and y = Excite



There will be no significant change in excitement level mainly because level of injury/illness is not specified so it could be extreme or nothing that impacts the dogs life. The dot plot does not really show much difference with healthy dogs having a slightly higher mean and first quartile. The health of a dog can effect the excitability in Labrador retrievers.

t.test with x = Health and y = Excite

```
##
##  Welch Two Sample t-test
```

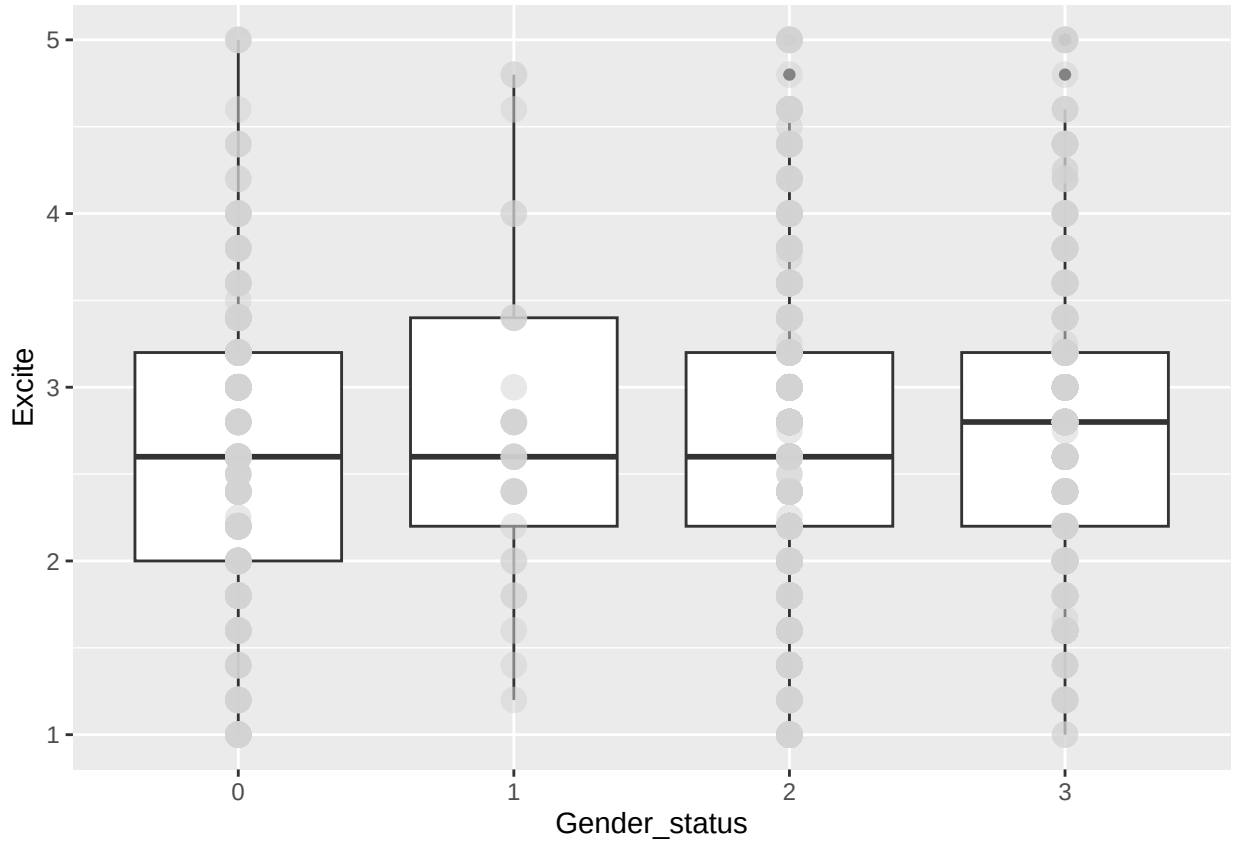
```
##
## data:  Excite by Health
## t = -1.2115, df = 377.75, p-value = 0.2265
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.17166393  0.04077547
## sample estimates:
## mean in group 0 mean in group 1
##        2.702245        2.767690
```

There is no significant difference between some heath issues and no health issues when it comes to the level of excitement of Labrador retrievers p = 0.2265. We reject the hypothesis. This is no real surprise as health problems vs no health problems is rather vague and because this comes from a survey it is hard to compare scale of injury.

Gender status vs Excitability

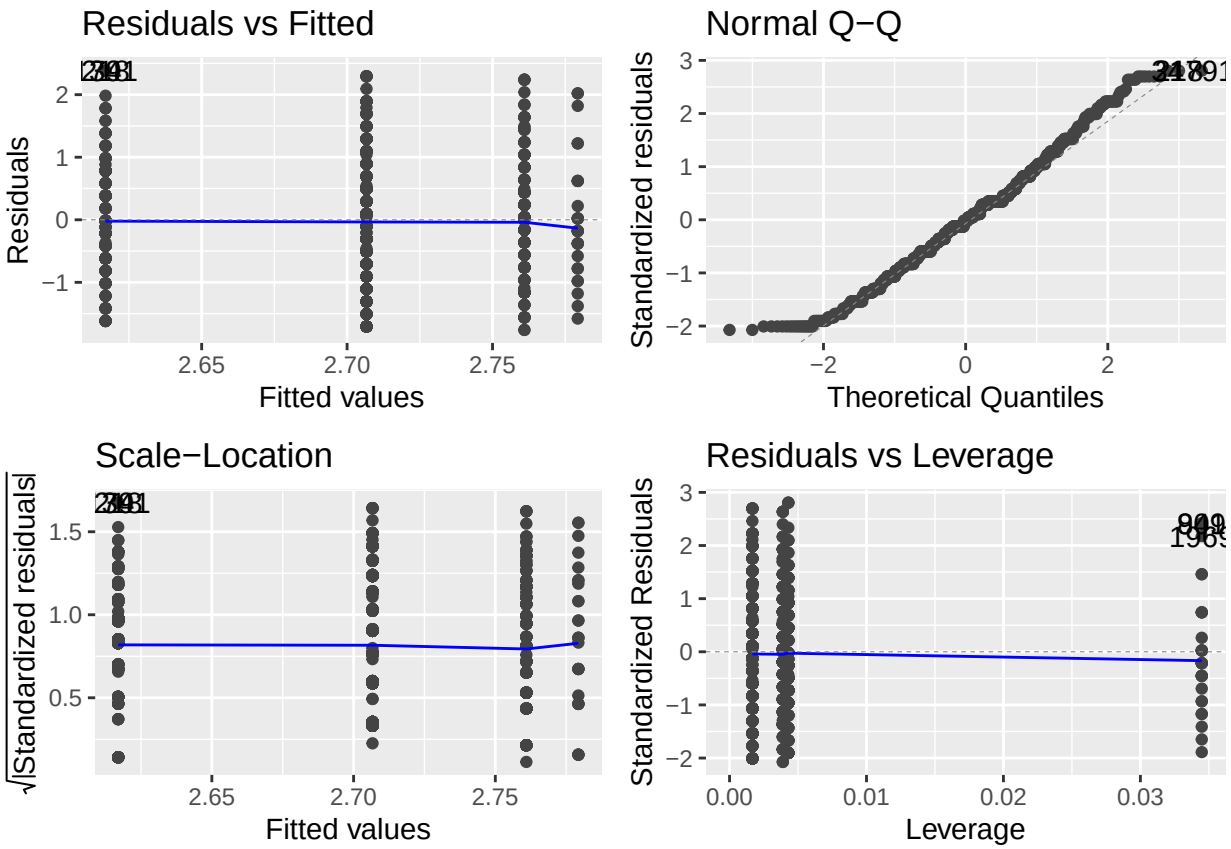
these last few are just for fun and to look at all the predictor variables so I wont go in deep. 0 (n = 451), 1 (n = 59), 2 (n = 1028), 3 (n = 426)

box plot of x = Gender_status and y = Excite



there does not seem to be any major differences between gender status with most means around 2.6 with the exception of fixed females where its closer to 2.8 but it has a similar IQR. There will be a difference in excitability in dogs between gender statuses.

autoplot of linear model with x = Gender_status and y = Excite




```
##
## Call:
## lm(formula = Excite ~ Gender_status, data = na.omit(DOG))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76096 -0.60754 -0.01695  0.49332  2.38305
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.61695    0.05576  46.934   <2e-16 ***
## Gender_status1  0.16236    0.16760   0.969   0.3329
## Gender_status2  0.08973    0.06571   1.365   0.1724
## Gender_status3  0.14401    0.07699   1.870   0.0617 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8511 on 1114 degrees of freedom
## Multiple R-squared:  0.003415,   Adjusted R-squared:  0.0007309
## F-statistic: 1.272 on 3 and 1114 DF,  p-value: 0.2825
```

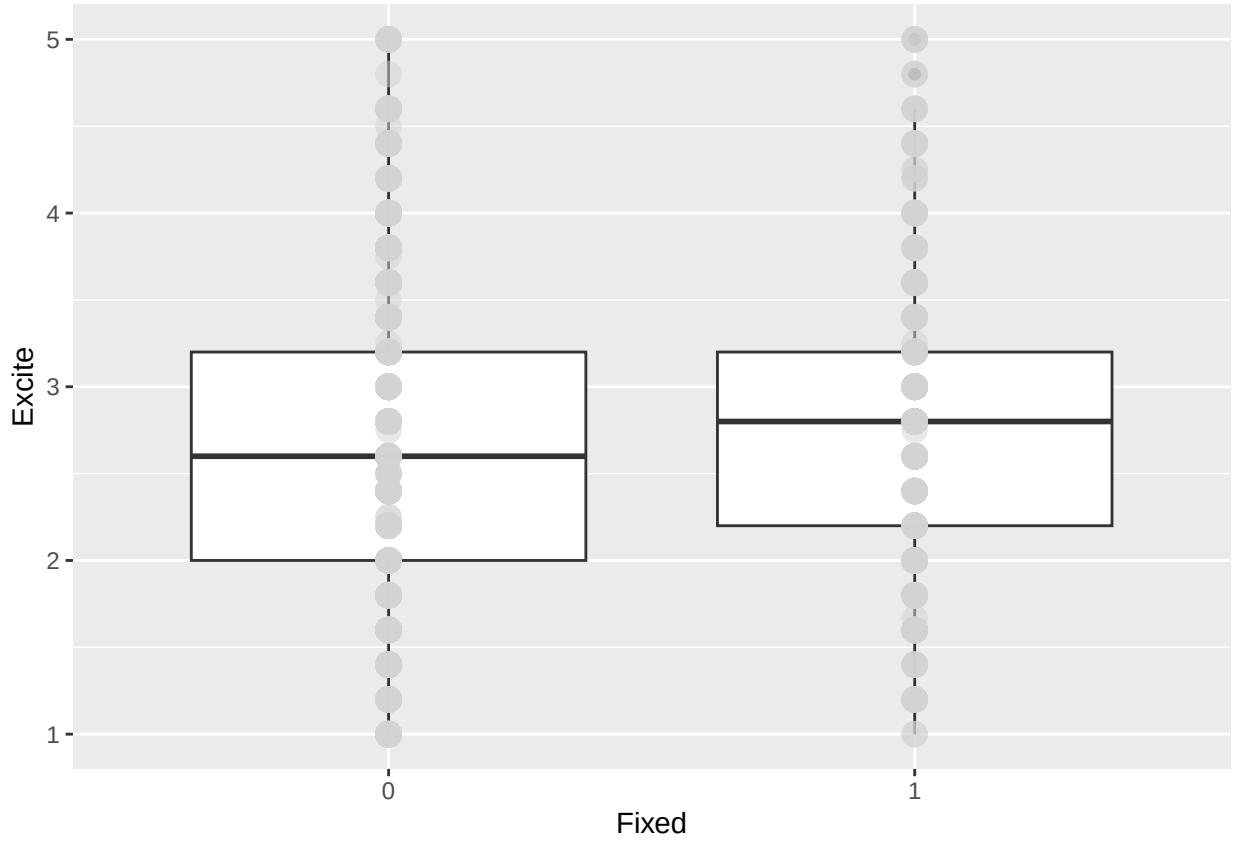
the autoplot looks good though there is a slight downward turn in both residual vs fitted and residual leverage for one column of dots which is likely fixed males because there are only 59 entries.

we reject the hypothesis as there is no significant difference between any of the values. with the closes to a significant value being p = 0.0617. The R-squared value is quite small at 0.003415.

Fixed vs Excitability

has the dog been Fixed. did the math near the top of the file to separate out the fixed and non-fixed dogs from Gender_status

box plot of x = Fixed and y = Excite



there does not seem to be a difference between fixed and non-fixed dogs when it comes to their excitability. Whether a dog is fixed or not will effect there excitability.

t.test with x = Fixed and y = Excite

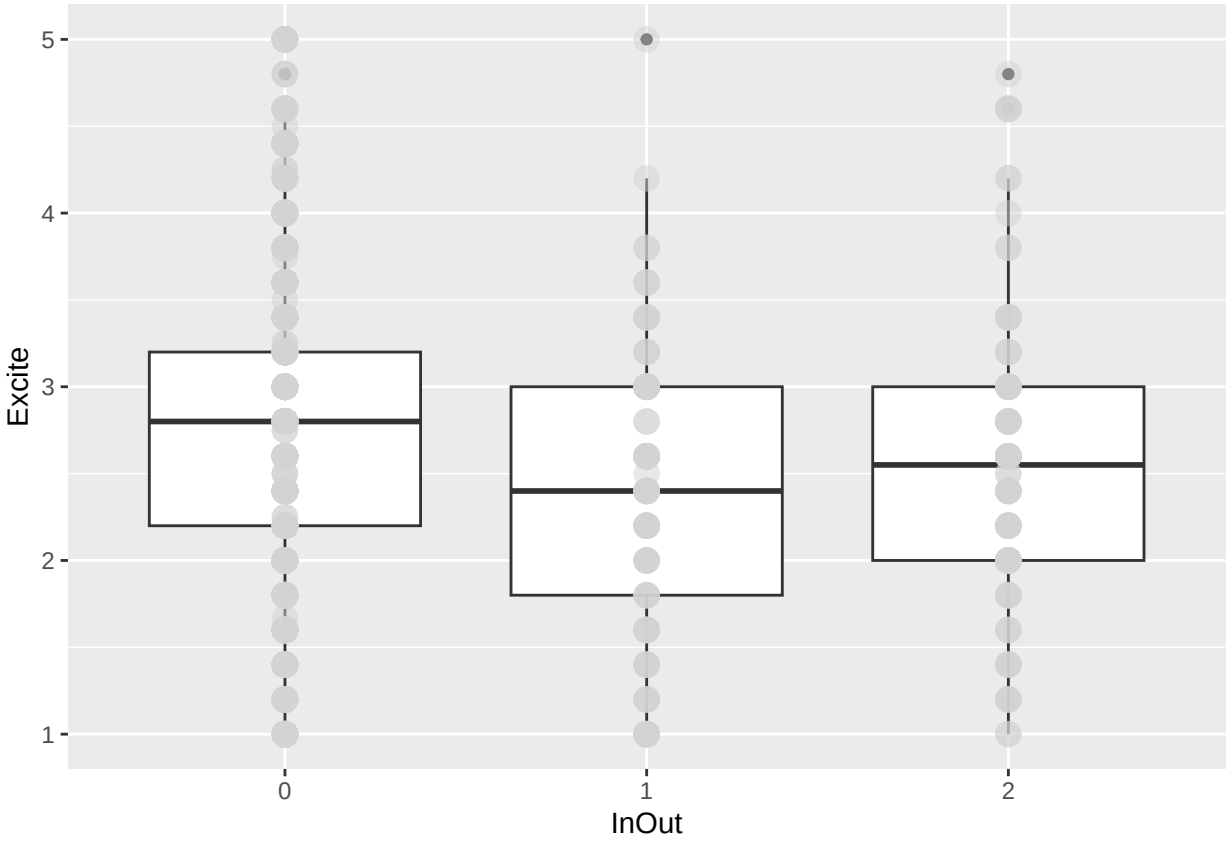
```
##
##  Welch Two Sample t-test
##
## data:  Excite by Fixed
## t = -0.75672, df = 832.21, p-value = 0.4494
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.12047178  0.05342856
## sample estimates:
## mean in group 0 mean in group 1
##        2.702670        2.736191
```

we reject the hypnosis that whether a dog is fixed or not effects its level of excitability with a p-value = 0.4494. it is very far from being significant.

InOut vs Excitability

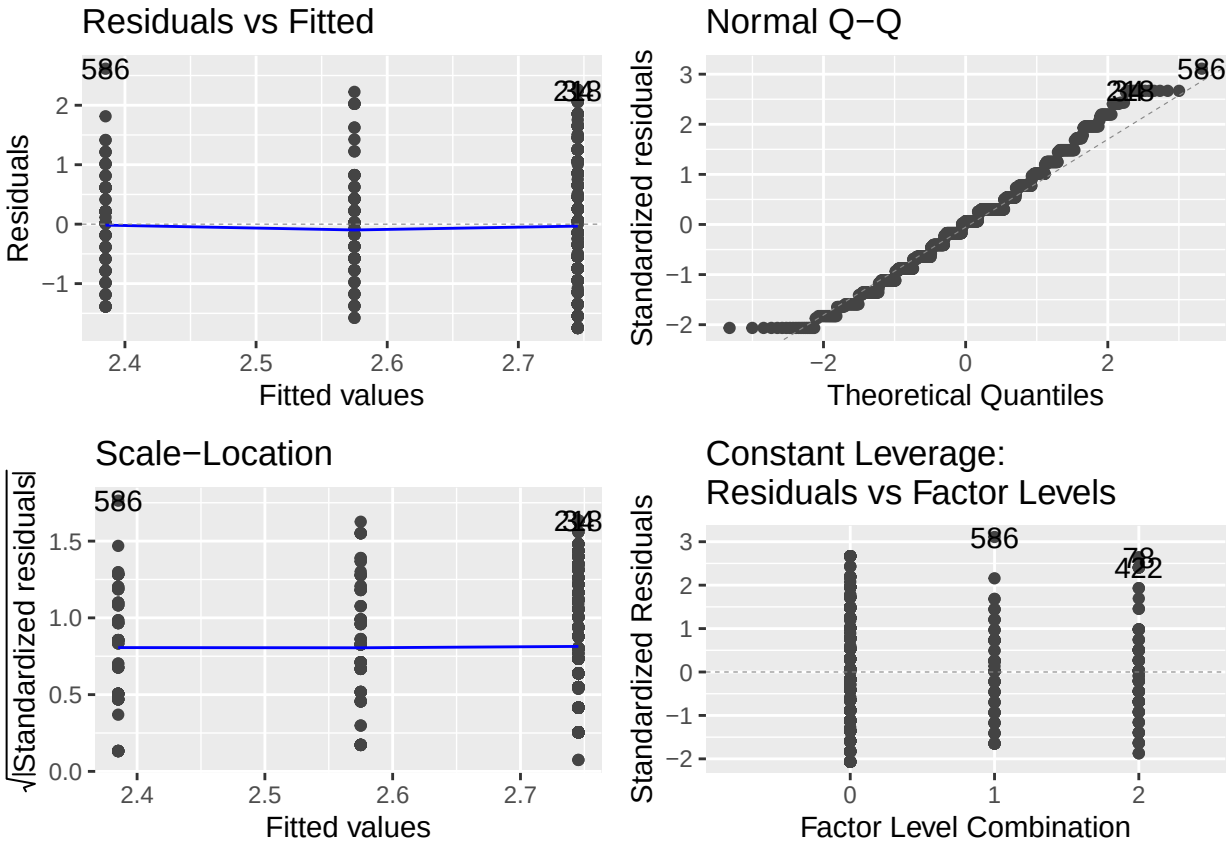
does where the dog live effect the level of excitability most of the dogs were indoor (n = 1578), indoor/outdoor (n = 170), outdoor (n = 176)

box plot of x = InOut and y = Excite



there does not seem to be a major difference, maybe indoor dogs are slightly more excitable. where a dog is housed effects its level of excitability.

autoplot of linear model with x = Fixed and y = Excite




```
##
## Call:
## lm(formula = Excite ~ InOut, data = na.omit(DOG))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74531 -0.54531  0.05469  0.45469  2.61477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74531    0.02767  99.199  < 2e-16 ***
## InOut1      -0.36008    0.09431  -3.818 0.000142 ***
## InOut2      -0.17031    0.09065  -1.879 0.060542 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8458 on 1115 degrees of freedom
## Multiple R-squared:  0.01498,    Adjusted R-squared:  0.01321
## F-statistic: 8.479 on 2 and 1115 DF,  p-value: 0.0002216
```

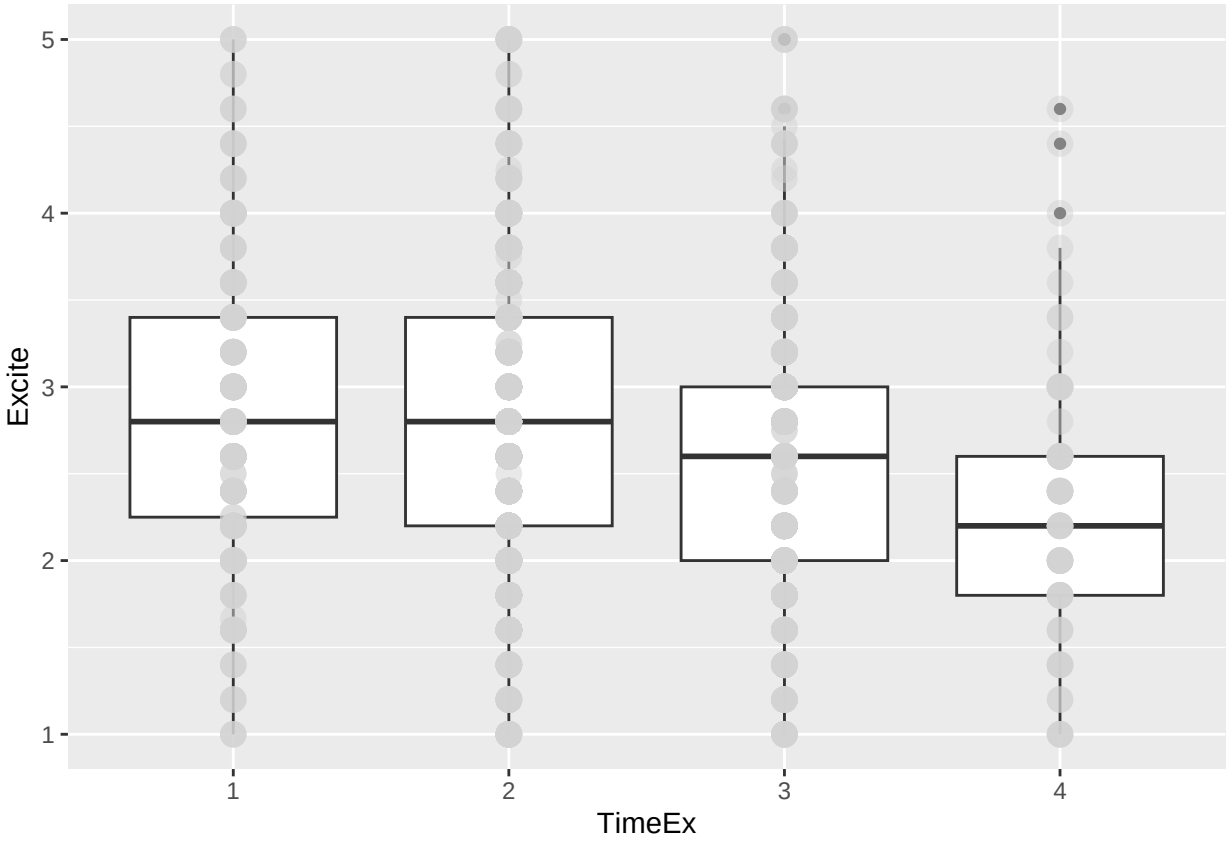
the autoplot seems fine with seemingly little issues or deviations off of the grey lines.

there is a significant difference between indoor dogs and indoor/outdoor dogs with a p-value = 0.000142 but there is no significant difference between indoor and outdoor dogs p-value = 00.060542. we reject the hypothesis that housing effects level of excitability due to the outdoor vs indoor comparison. Is there a difference in the job of indoor/outdoor dogs vs outdoor/are owners of gundogs more likely to leave their dogs outside? this was a very unexpected result to even have a significant value.

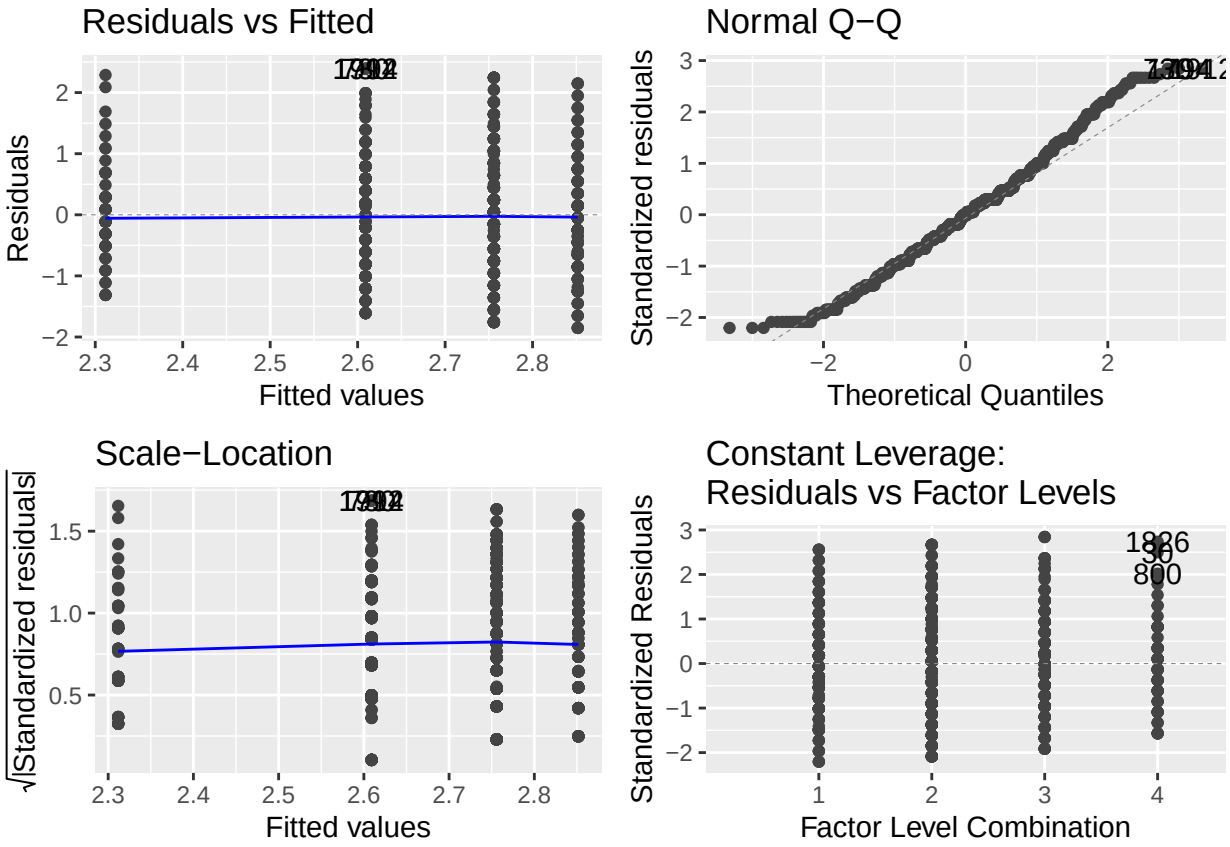
Time exercising vs Excitability

does the amount of hours exercising effect the excitability of dogs?

box plot of x = TimeEx and y = Excite



there seems to be a negative correlation between exercise and excitability. How much a dog exercises effects its level of excitability.




```
##
## Call:
## lm(formula = Excite ~ TimeEx, data = na.omit(DOG))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85168 -0.55578 -0.00908  0.44422  2.39092
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.85168    0.06129  46.531  < 2e-16 ***
## TimeEx2     -0.09589    0.07128  -1.345  0.17879
## TimeEx3     -0.24260    0.07707  -3.148  0.00169 **
## TimeEx4     -0.53991    0.11914  -4.532 6.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8425 on 1114 degrees of freedom
## Multiple R-squared:  0.0234, Adjusted R-squared:  0.02077
## F-statistic: 8.896 on 3 and 1114 DF,  p-value: 7.916e-06
```

the autoplot looks very good with no real issues that can be seen or that are majorly different in terms of Normal Q-Q.

there is no significant difference between 1 and 2 hours (p-value = 0.17879) but there are significant differences between 1 and 3 (p-value = 0.00169) and 1 and 4 hours (p-value = 6.49e-06). We reject the hypothesis that there is a difference between amount of time exercising and excitability due to the p-value between 1 and 2

hrs. It seems as dogs exercise they get tired out and less excitable but that can not be confirmed with this test but is an interesting result.

Biological Summary

looking only at a Labrador retrievers excitability and comparing it to all seven predictor variables we found that only job had a significant impact on excitability. Although a few other of the predictors did have significant values for part of the tests there were also parts of the same predictor that were not significant so we could not confirm the hypotheses but under different tests or hypotheses they could show differences. For the paper that we got the data from, Joanna Ilska, et al., they also found role/job to be statistically significant but they also found exercise and housing to be significant. From the data analysis there is no clear impact of a dogs nature but there is evidence that how a dog is nurtured, at least in the way of its job, will effect its excitability.

Challenges

The main challenges for me were to figure out what how to manipulate the data that I am using because there is only 1 set of data that is not between 1-5 in this data set, this made it difficult to figure out what test to use as a lot of data was semi-continuous but based off a questionnaire. It was interesting to learn how difficult questionnaires can be to manipulate data from because they are used extensively outside of biology, and often in biology, to get data. One problem that I had was that I initially was confused about what test I should be using and there is still some evidence of that in my data analysis. Something cool that I did was use the mod function to separate out whether a dog had been fixed or not from gender status, though I am not sure if I could get actual gender out of the data. Overall I learned that playing with data is fun but it can be time consuming and the data might be hard to understand and manipulate.