Astronaut Data

Adam Gruber

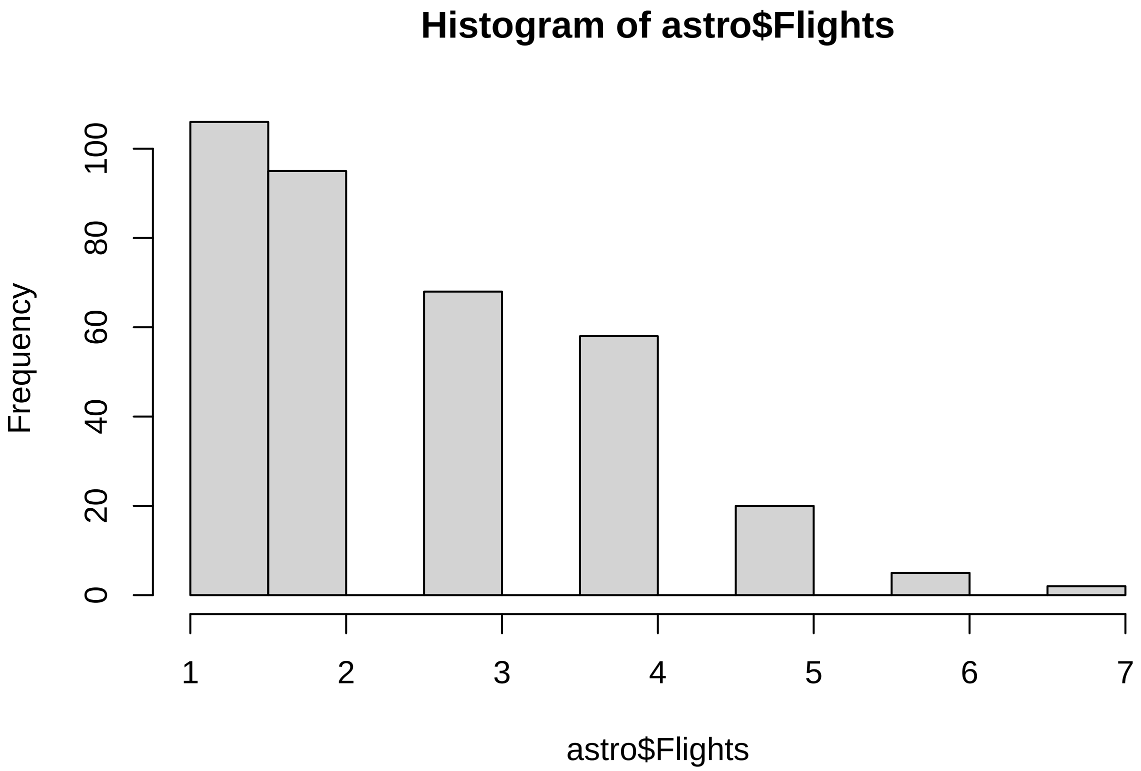
2023-09-29

R Markdown

Initially when exploring the data there appears to be several NAs in the data. One of the NAs is for Children. I noticed a lack of zero in the data. So i converted the NA to zero. The time for the astronauts in space is broken up into 3 columns: Days, Minutes, Hours.

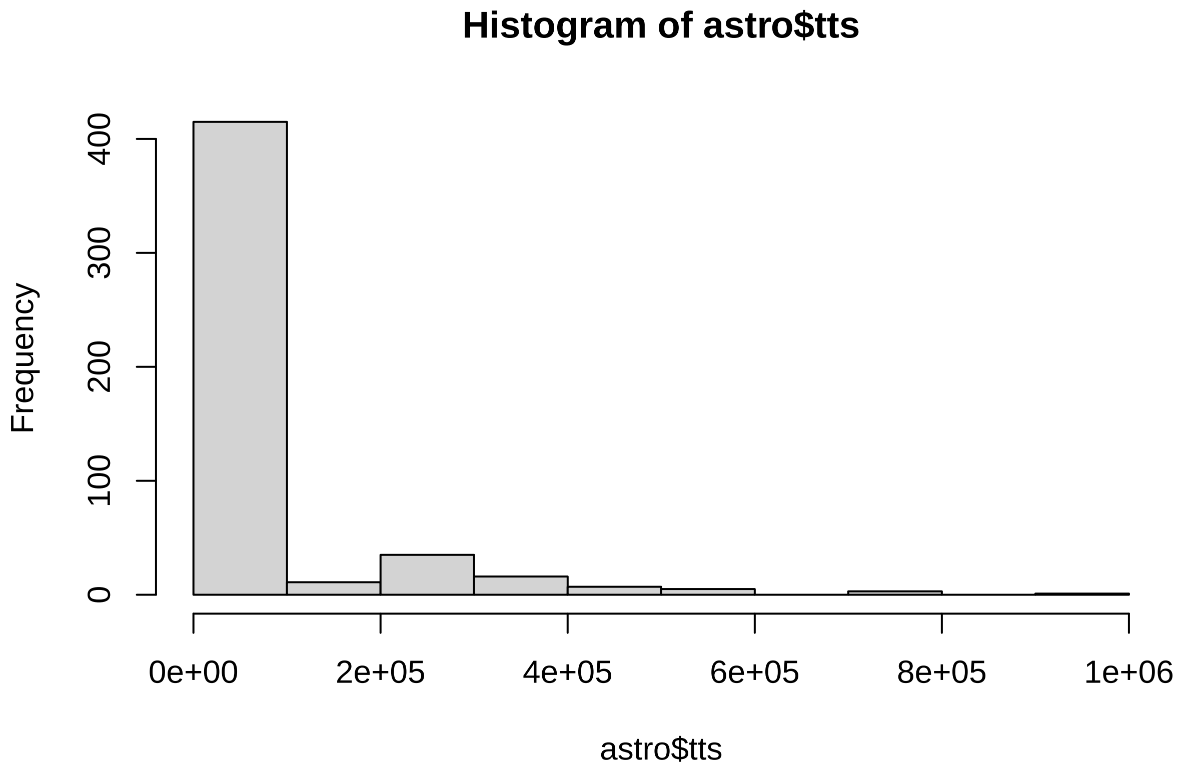
```
##    Surname           Given.names        Date.of.Birth      Place.of.Birth    
##  Length:493         Length:493         Length:493         Length:493        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##   Selection           Position            Status             Flights     
##  Length:493         Length:493         Length:493         Min.   :1.000  
##  Class :character   Class :character   Class :character   1st Qu.:1.000  
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000  
##                                                           Mean   :2.475  
##                                                           3rd Qu.:3.000  
##                                                           Max.   :7.000  
##                                                           NA’s   :139    
##        d                h                m         Marital.status
##  Min.   :  0.00   Min.   : 0.000   Min.   : 0.0   divorced: 22   
##  1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.: 0.0   married :420   
##  Median : 15.00   Median : 5.000   Median :15.0   single  : 20   
##  Mean   : 47.87   Mean   : 8.258   Mean   :21.3   widowed :  8   
##  3rd Qu.: 40.00   3rd Qu.:16.000   3rd Qu.:41.0   NA’s    : 23   
##  Max.   :665.00   Max.   :23.000   Max.   :59.0                  
##                                                                  
##     Children        Nation          Date.of.Death      Place.of.Death    
##  Min.   :0.000   Length:493         Length:493         Length:493        
##  1st Qu.:2.000   Class :character   Class :character   Class :character  
##  Median :7.000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :5.016                                                           
##  3rd Qu.:8.000                                                           
##  Max.   :8.000                                                           
## 
```

The column fights is right skewed. There are a few exceptional people that have done more than just 1 or 2

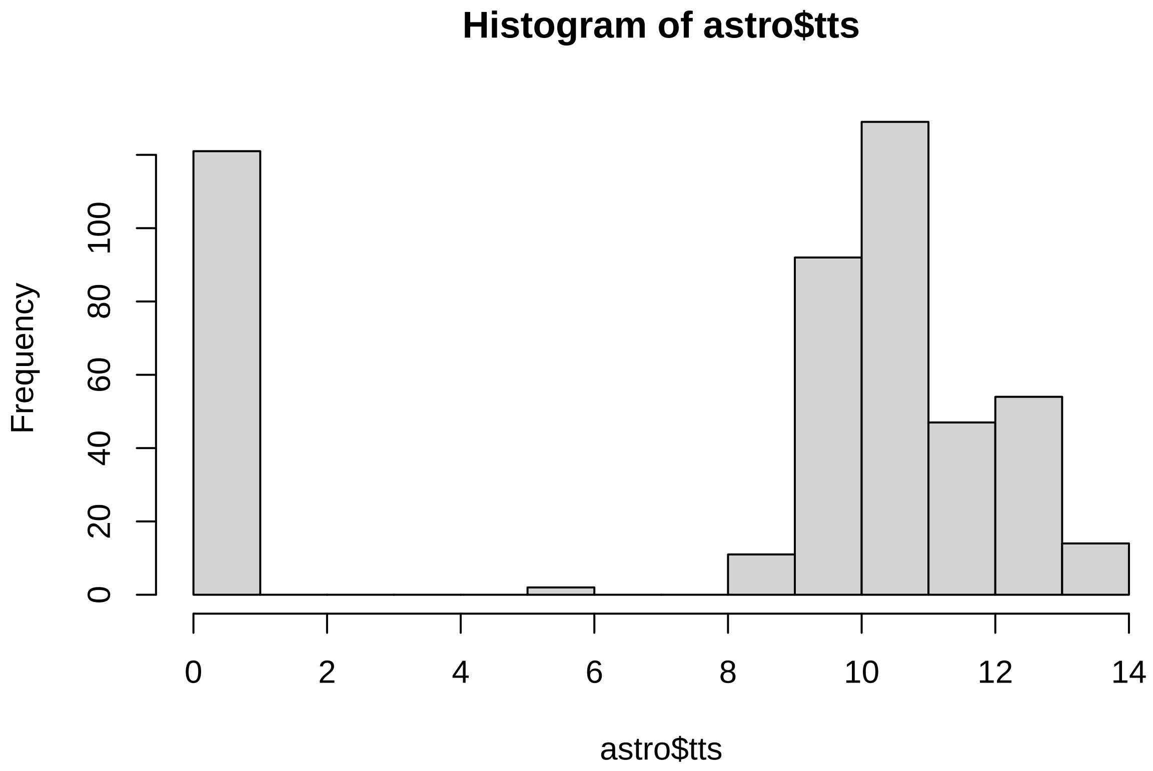


flights.

To better calculate the data I created a new column called Total Time in Space(TTS). This allows us to have an accurate comparison. This column is still skewed to the right as well. The majority of astronauts having less than 1000 hours in space. The upper levels would be considered outliers. They are astronauts that spent the most time on the Interantional Space Station. Keeping them in could be valuable as they spent more time away from loved ones.



I removed the N/A rows from Marital Status to ensure proper data. I also took all the flights that were NA and converted them to 0



I tested several different models for a fit with the astronaut data trying to determine if time in space and or flights affected the odds of getting divorced. I took the factor column of Marital Status and added a new column Mar that was

The first was the linear model. All of the predictors are considered not significant. All of the linear models have really correlation. Less than .1. The model is little more than random guessing.

```
##
## Call:
## lm(formula = mar ~ tts + Flights, data = astro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04307 -0.03606 -0.03175 -0.01733  1.98267
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.0173263  0.0354557  56.897   <2e-16 ***
## tts         0.0007678  0.0056567   0.136    0.892
## Flights     0.0034305  0.0171396   0.200    0.841
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3969 on 467 degrees of freedom
## Multiple R-squared:  0.0004653,  Adjusted R-squared:  -0.003815
## F-statistic: 0.1087 on 2 and 467 DF,  p-value: 0.897
```

```
##
## Call:
## lm(formula = mar ~ tts, data = astro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03722 -0.03485 -0.03323 -0.01702  1.98298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.017017   0.035386  57.001   <2e-16 ***
## tts         0.001605   0.003806   0.422    0.674
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3964 on 468 degrees of freedom
## Multiple R-squared:  0.0003796,  Adjusted R-squared:  -0.001756
## F-statistic: 0.1777 on 1 and 468 DF,  p-value: 0.6735
```

```
##
## Call:
## lm(formula = mar ~ tts + Flights + Children, data = astro)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07666 -0.04363 -0.02350 -0.01458  1.99595
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.045895   0.045705  44.763   <2e-16 ***
## tts          0.001036   0.005663   0.183    0.855
## Flights      0.003832   0.017145   0.224    0.823
## Children    -0.005977   0.006034  -0.991    0.322
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3969 on 466 degrees of freedom
## Multiple R-squared:  0.002566,   Adjusted R-squared:  -0.003856
## F-statistic: 0.3995 on 3 and 466 DF,  p-value: 0.7534
```

The second model used for making prediction was the logistic model using the glm function. None of the predictors are considered significant. The AIC between all 3 models is relatively close with 3 points: 470,468,471.

```
##
## Call:
## glm(formula = mar ~ tts + Flights, data = astro)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.0173263  0.0354557  56.897   <2e-16 ***
## tts         0.0007678  0.0056567   0.136    0.892
## Flights     0.0034305  0.0171396   0.200    0.841
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.1574919)
##
##     Null deviance: 73.583  on 469  degrees of freedom
## Residual deviance: 73.549  on 467  degrees of freedom
## AIC: 470.05
##
## Number of Fisher Scoring iterations: 2
```

```
##
## Call:
## glm(formula = mar ~ tts, data = astro)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.017017   0.035386  57.001   <2e-16 ***
## tts         0.001605   0.003806   0.422    0.674
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.1571689)
##
##     Null deviance: 73.583  on 469  degrees of freedom
## Residual deviance: 73.555  on 468  degrees of freedom
## AIC: 468.09
##
## Number of Fisher Scoring iterations: 2
```

```
##
## Call:
## glm(formula = mar ~ tts + Flights + Children, data = astro)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.045895   0.045705  44.763   <2e-16 ***
## tts          0.001036   0.005663   0.183    0.855
## Flights      0.003832   0.017145   0.224    0.823
## Children    -0.005977   0.006034  -0.991    0.322
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.1574983)
##
##     Null deviance: 73.583  on 469  degrees of freedom
## Residual deviance: 73.394  on 466  degrees of freedom
## AIC: 471.06
##
## Number of Fisher Scoring iterations: 2
```

We try the LDA model to make predictions on who will get divorced based on time in space and flights. This model accurately classifies 420/470. It classifies all astronauts as married. They made up the largest group and the model tries to predict as accurately as possible. it is only 89% accurate.

```
ldafit1 = lda(mar~tts+Flights, data = astro)


ldafit1
```

```
## Call:
## lda(mar ~ tts + Flights, data = astro)
##
## Prior probabilities of groups:
##          1          2          3          4
## 0.04680851 0.89361702 0.04255319 0.01702128
##
## Group means:
##        tts  Flights
## 1 7.623548 1.363636
## 2 7.963446 1.883333
## 3 7.998324 1.800000
## 4 8.536300 1.625000
##
## Coefficients of linear discriminants:
##                LD1         LD2
## tts      0.2158797 -0.22033367
## Flights -0.9352442  0.05486455
##
## Proportion of trace:
##    LD1    LD2
## 0.9564 0.0436
```

```
fittedclasslda1 = predict(ldafit1,data = astro)$class
table(astro$mar,fittedclasslda1)
```

```
##    fittedclasslda1
##       1   2   3   4
##   1   0  22   0   0
##   2   0 420   0   0
##   3   0  20   0   0
##   4   0   8   0   0
```

I tried a simpler model that only uses the number of Minutes in space. This produced the same accuracy as before when predicting.

```
ldafit2 = lda(mar~tts, data = astro)
ldafit2
```

```
## Call:
## lda(mar ~ tts, data = astro)
##
## Prior probabilities of groups:
##          1          2          3          4
## 0.04680851 0.89361702 0.04255319 0.01702128
##
## Group means:
```

```
##        tts
## 1 7.623548
## 2 7.963446
## 3 7.998324
## 4 8.536300
##
## Coefficients of linear discriminants:
##          LD1
## tts 0.207313
```

```
fittedclasslda2 = predict(ldafit2,data = astro)$class
table(astro$mar,fittedclasslda2)
```

```
##    fittedclasslda2
##       1   2   3   4
##   1   0  22   0   0
##   2   0 420   0   0
##   3   0  20   0   0
##   4   0   8   0   0
```

LDA model 3 adds in children.

```
ldafit3 = lda(mar~tts+Flights + Children, data = astro)


ldafit3
```

```
## Call:
## lda(mar ~ tts + Flights + Children, data = astro)
##
## Prior probabilities of groups:
##          1          2          3          4
## 0.04680851 0.89361702 0.04255319 0.01702128
##
## Group means:
##        tts  Flights Children
## 1 7.623548 1.363636 2.590909
## 2 7.963446 1.883333 5.623810
## 3 7.998324 1.800000 0.100000
## 4 8.536300 1.625000 6.500000
##
## Coefficients of linear discriminants:
##                  LD1         LD2          LD3
## tts      -0.04219109 -0.21987575  0.213190842
## Flights   0.09980640  0.93123683 -0.023231045
## Children  0.35887268 -0.03988201 -0.005021991
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.9653 0.0328 0.0019
```

```
fittedclasslda3 = predict(ldafit3,data = astro)$class
table(astro$mar,fittedclasslda3)
```

```
##    fittedclasslda3
##       1   2   3   4
##   1   0  22   0   0
##   2   0 420   0   0
##   3   0  20   0   0
##   4   0   8   0   0
```

Here I try the QDA model to properly classify the astronaut data based on their marital status. Model 1 is tts and FLights Model 2 is tts Model 3 is tts Flights, Children

```
qdafit1 = qda(mar~ tts + Flights, data = astro)
astro%>%
  group_by(mar)%>%
  summarize(sdtts = sd(tts), sdFlights = sd(Flights))
```

```
## # A tibble: 4 x 3
##     mar sdtts sdFlights
##   <int> <dbl>     <dbl>
## 1     1  4.44      1.43
## 2     2  4.86      1.60
## 3     3  4.81      1.64
## 4     4  3.51      1.19
```

```
fittedclassqda1 = predict(qdafit1,data = astro)$class
table(astro$mar, fittedclassqda1)
```

```
##    fittedclassqda1
##       1   2   3   4
##   1   0  22   0   0
##   2   0 420   0   0
##   3   0  20   0   0
##   4   0   8   0   0
```

```
diag(table(astro$mar,fittedclassqda1))
```

```
##   1   2   3   4
##   0 420   0   0
```

```
qdafit2 = qda(mar~ tts, data = astro)

fittedclassqda2 = predict(qdafit2,data = astro)$class
table(astro$mar, fittedclassqda2)
```

```
##    fittedclassqda2
##       1   2   3   4
##   1   0  22   0   0
##   2   0 420   0   0
##   3   0  20   0   0
##   4   0   8   0   0
```

```
diag(table(astro$mar,fittedclassqda2))
```

```
##   1   2   3   4
##   0 420   0   0
```

The QDA when using the the predictors: Children, Flights, TTS, actually does worse. Adding in too many variables seems to have reduced accuracy. It correctly predicted 405/470. That is 86% accuracy. This is likely due to the model better predicting the factor level Single. Single people are less likely to have children. The model previously did not cassify anyone as single when trying to classify them. Now it correctly classified 15/20 people in the data set.

```
qdafit3 = qda(mar~ tts + Flights +Children, data = astro)


fittedclassqda3 = predict(qdafit3,data = astro)$class
table(astro$mar, fittedclassqda3)
```

```
##    fittedclassqda3
##       1   2   3   4
##   1   0  13   9   0
##   2   0 390  30   0
##   3   0   5  15   0
##   4   0   8   0   0
```

```
diag(table(astro$mar,fittedclassqda3))
```

```
##   1   2   3   4
##   0 390  15   0
```

Using the higher accuracy QDA model, I used 10 fold cross validation to test and see what level of accuracy we can predict the marital status. The real goal is to predict the likely hood of divorce.

```
## [1] "Model 1 Error"
```

```
## [1] 0.106383
```

```
## [1] "Model 2 Error"
```

```
## [1] 0.106383
```

The model shows a low cross validation error of .106 This would normally be good but in our data set, we still do not accurately predict the divorces that will take place. We are merely accurately predicting that married people will continue to be married.

We did not build a model to accurately predict divorces among astronauts based on their time in space. We did however find that astronauts are much more likely to stay together compared to the normal population. The normal population experiences a divorce rate of 40 -50%. The astronaut population experiences 5.2 % divorce rate. This small portion of the overall population makes it hard for the models to predict accurately when the person will get divorced.

The hypothesis was stress from being an astronaut and time in space would lead to a divorce.

We reject the null hypothesis. Time in space is not significantly correlated with divorce.

We have enough evidence to prove the disprove the common myth that astronauts have a higher divorce rate due to stress of the job and time spent in sace.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.