Capstone Data Science PH125.9x: Reality TV, the Bachelor

Hema Murty

2024-06-11

INTRODUCTION

This project looks at the data from the Reality TV series called The Bachelor (Wiki The Bachelor, n.d.).

On first look, it seemed that after a long run of the series, there surely must be enough data to put together a viable data science analysis of the show's end results and some variables that might affect those results.

However, once I took the data available on the internet, from various data sources, the data, even after 28 seasons of the show, was still too low for a conclusive analysis.

We would need at least another 28 seasons of the show's run to have enough data that might generate faith in the analysis.

Therefore, since the goal of the capstone is to use techniques that we learned in the course on data that is available publicly and make conclusions based on that data, I decided to see what conclusions I could make.

Also, there was a timeline issue that prevented me from completing this section earlier, owing to family members passing.

Therefore, I used this dataset to complete the project, rather than exploring other subject matter and their corresponding datasets which would have taken longer amidst a looming deadline to submit.

America has a fascination with the longest running reality television dating show. Its most recent season 28 had approximately 3 million viewers.

In data science, we always want to know if there are some determinants for the results and what part they would play in predicting future results.

Surprisingly there have been many studies on the determining factors of the results. This is owing to online betting sites for this show. (Entertainment Betting/ The Bachelor, n.d.) Society is obsessed with predicting the winner of each season.

In this project, I chose to look at the distance between the participant's hometown to the lead Bachelor's hometown as a deciding factor for their elimination week.

This report shows the results obtained and the conclusion from that query.

METHODS

There is a lot of publicly available data on The Bachelor on Kaggle and data.world. Indeed, Wikipedia has information on each season's placement of participants.

The data was cleaned up to focus on United States participants to make distance calculations using Google's API on Geography amenable to the data science analysis.

Therefore, international contestants were removed from the dataset, noting that none of the international participants have ever advanced to winners in the show's history.

There was one bachelor lead who was from Canada and that season did not have a clear winner. So that season was eliminated from the dataset.

The most difficult part of this project was cleaning the data. Owing to the amount of work in cleaning the data and that data for all 28 seasons of the show is found in different datasets in different locations on the internet in different formats, the looming deadline did not permit me to include as many of the 28 seasons as I would have liked for the training.

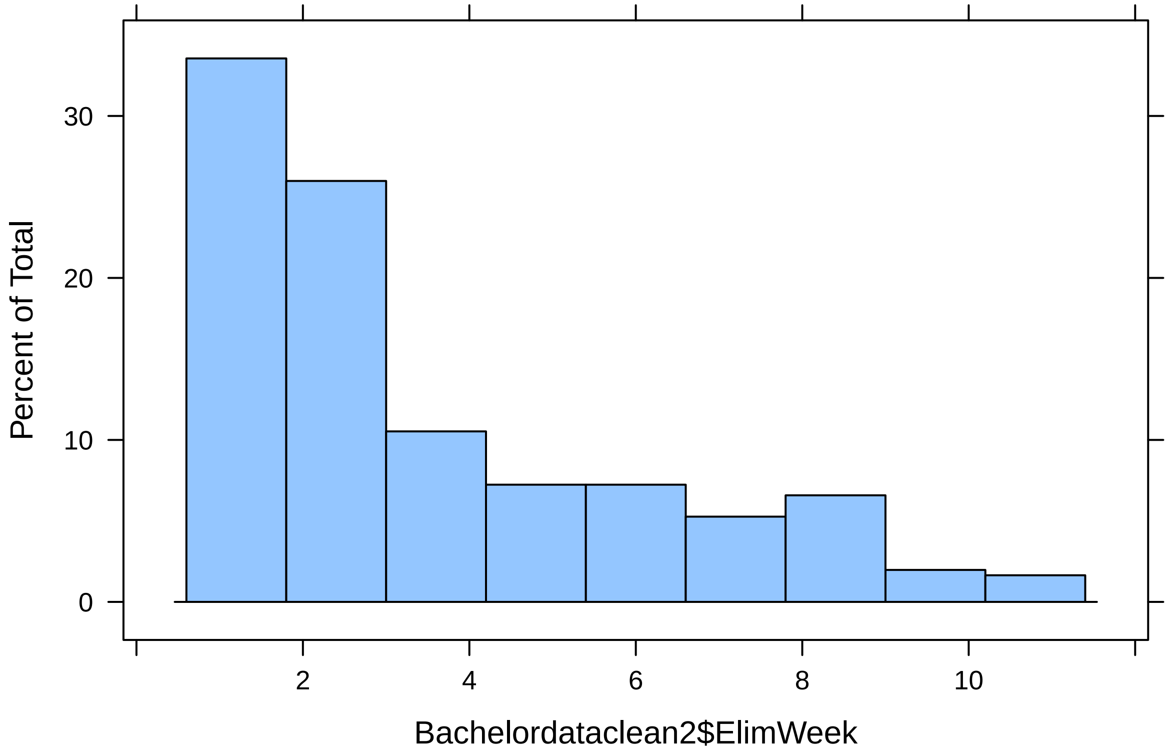
I used 20 seasons of the show. Data wrangling for this project proved time consuming with data scattered in various formats across the internet. Only one technical paper, published in 2022, looked at a data analysis investigation of this television show from different seasons than those that I used here and their methods of analysis were different than the ones that I used here. (AJ Lee, 2022)

I did the cleaning of the dataset and the calculation of distances in Excel prior to loading it into Rstudio because I needed a Google API to calculate the relative distance between the lead and participant's hometowns. I wanted that API to stay private so I had to pre-process the data and create a dataset which had that calculated distances to load into Rstudio in order to provide that as a deliverable for this project.

In addition, when creating training and testing datasets, proper names could not be included as the analysis would say that the test data did not have any overlap for the proper names info columns, with the training dataset. Therefore, a reduced dataset was created for analysis with just the numerical data.

First, we load the libraries required.

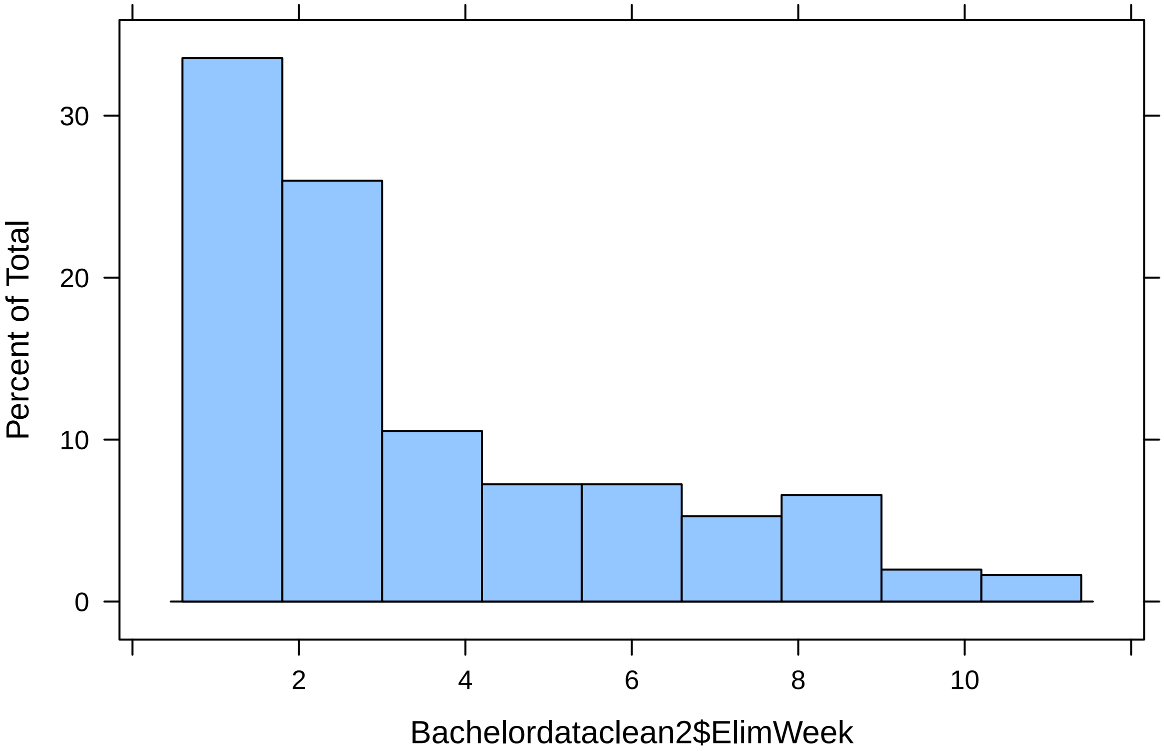
```
library(readxl)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(caret)
library(gridExtra)
library(dslabs)

#Read the data file from folder
#download data files Bachelordataclean.xlsx and Elimdistance.xlsx and place them in a subdirectory call

Bachelordataclean <- read_xlsx("./data/Bachelordataclean.xlsx")

#this is not a data frame but incorporates a tibble

Bachelordataclean2 <- as.data.frame(Bachelordataclean)
#take a look at the elimination weeks

histogram(Bachelordataclean2$ElimWeek)
```



```
summary(Bachelordataclean2)
```

```
##      Name                Age        Participant Hometown
##  Length:304         Min.   :21.00   Length:304
##  Class :character   1st Qu.:24.00   Class :character
##  Mode  :character   Median :26.00   Mode  :character
##                     Mean   :26.47
##                     3rd Qu.:28.00
##                     Max.   :36.00
##  Participant Hometown State Participant Hometown State2
##  Length:304                 Length:304
##  Class :character           Class :character
##  Mode  :character           Mode  :character
##
##
##
##  Participant Hometown State Abbrev     Season        Bachelor
##  Length:304                        Min.   : 1.00   Length:304
##  Class :character                  1st Qu.:11.00   Class :character
##  Mode  :character                  Median :15.00   Mode  :character
##                                    Mean   :13.14
##                                    3rd Qu.:18.00
##                                    Max.   :20.00
##   Bachelor Age   Bachelor_Hometown  Bachelor Home State
##  Min.   :28.00   Length:304         Length:304
##  1st Qu.:30.00   Class :character   Class :character
##  Median :32.00   Mode  :character   Mode  :character
##  Mean   :31.72
##  3rd Qu.:33.00
##  Max.   :35.00
##  Bachelor Hometown State Abbrev Bachelor Hometown State Participant’s Hometown
##  Length:304                     Length:304              Length:304
##  Class :character               Class :character        Class :character
##  Mode  :character               Mode  :character        Mode  :character
##
##
##
##  Bachelor Hometown     ElimWeek         Distance
##  Length:304         Min.   : 1.000   Min.   :   0.0
##  Class :character   1st Qu.: 1.000   1st Qu.: 650.2
##  Mode  :character   Median : 3.000   Median :1045.6
##                     Mean   : 3.493   Mean   :1104.5
##                     3rd Qu.: 5.000   3rd Qu.:1472.3
##                     Max.   :11.000   Max.   :4763.1
```

Including Plots



We can see that many participants have lower elimination weeks. This means that many participants are eliminated early leaving a fewer number of participants remaining as the show progresses.

A summary of all the data in that final dataset gives us the following information:

The minimum distance from the lead's hometown to the participant's hometown is 0 (meaning that the participant lived in the same town as the lead).

The maximum distance was 4763.1 miles.

The mean was 1104.5 miles amongst all seasons used in this project.

Let's discuss the spread in the distance data from the previous information of the summary.

Showing the mean distance to be 1045.6 miles between the lead's hometown and the participant's hometown. The standard deviation for the distance is 656.0478 miles. Considering the median is 1045.6 miles, this is about half of that distance.

There is a huge spread in distance of the participants.

The minimum Bachelor age was 28 years.

The maximum was 35 years and the mean was 31.72 years.

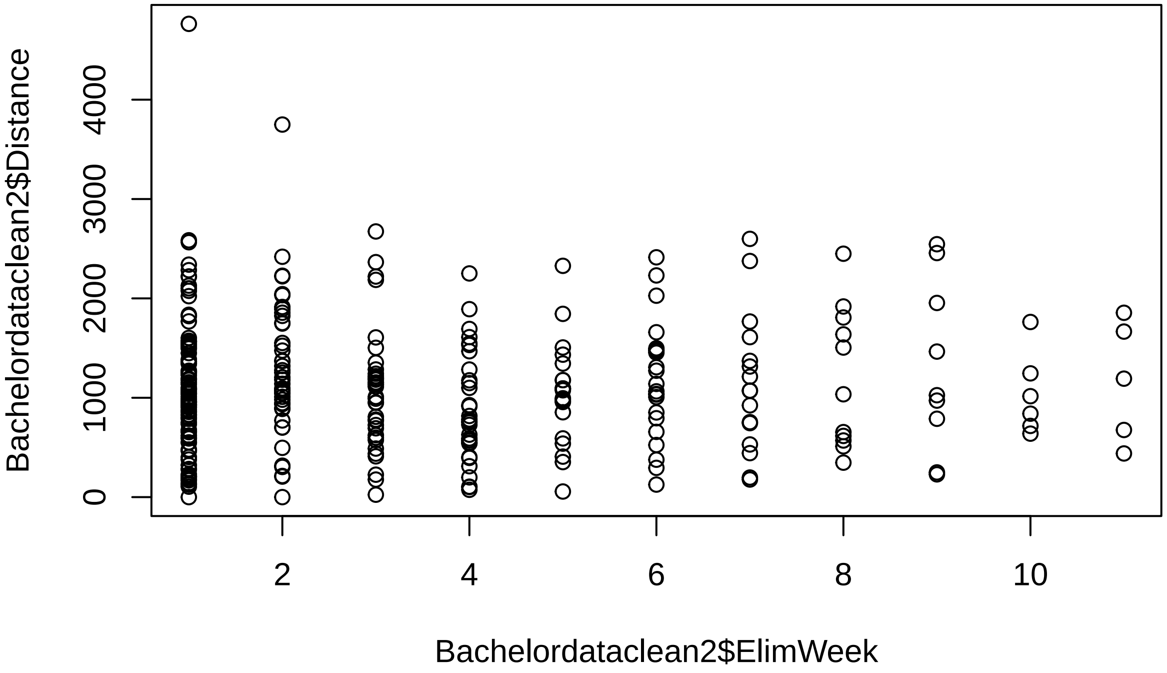
The minimum elimination week was 1. That is obvious from the premise of the show in which many participants are eliminated in week 1.

Of course, the maximum elimination week was 11 at the end of the season, there are two participants and one is sent home.

There are no NA data entries in the post-processed data.

The histogram of elimination week shows the progress of how many participants are sent home each week. Clearly, the first week has the maximum number of eliminations and as the show progresses each week, fewer participants are sent home.

We first look at a simple plot of the distance versus the elimination week to see if there is any kind of a relationship as a simple graph.



The plot showing distance versus the elimination week does not show that participants that have a closer distance to the lead's hometown will stay longer in the game.

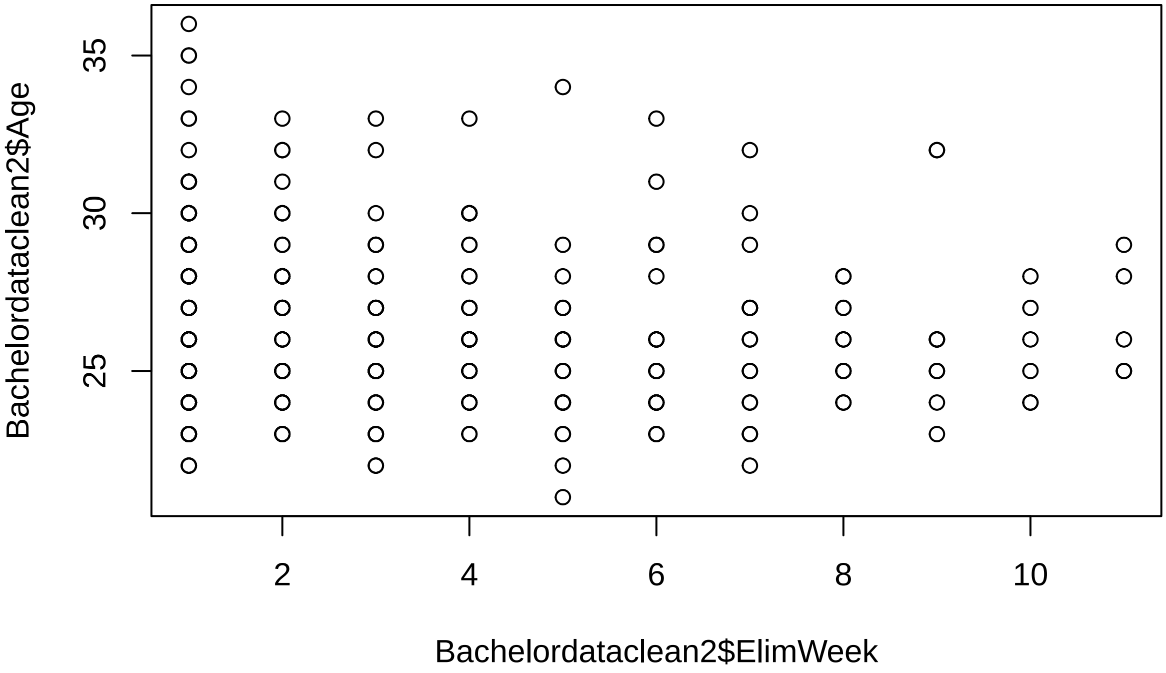
It seems to indicate that there is not a strong bias towards distance.

If I had more time, I would dive deeper into gathering other parameters that might affect the outcome.

This might be personality types, perhaps the career path or other non-traditional variables, such as the type of interactions that the bachelor lead had with each participant and the quality of that interaction.

Again, as the data is scattered in various formats and locations across the internet, this data cleaning for a deeper dive would take a few more weeks, as it might involve watching each episode for this kind of data.

Next, we look at the age of the participants to see if that has any bearing on the elimination week.



Again this data has a wide spread. It only shows that older participants are eliminated in week 1, indicating a bias towards participants under 30 years old.

Out of interest, the minimum age of the participants is 21 and maximum age is 36 with a median of 26 throughout the first twenty seasons. Standard deviation of the age is 2.759724 and so really close around the mean.

Next we look at splitting the dataset into a training and testing set with 70% as training set and 30% as test set.

Before we do this from the original data set, we must be aware that some columns have character names as entries for the participants. This will cause a problem if we split the data and expect the test data to have overlap with the training data.

Therefore, we will create a new dataset which is a subset of the larger dataset with just numerical data.

The training set has a dimension of 213 x 18 where we included a column “id” for convenience. The test set has a dimension of 91 x 18.

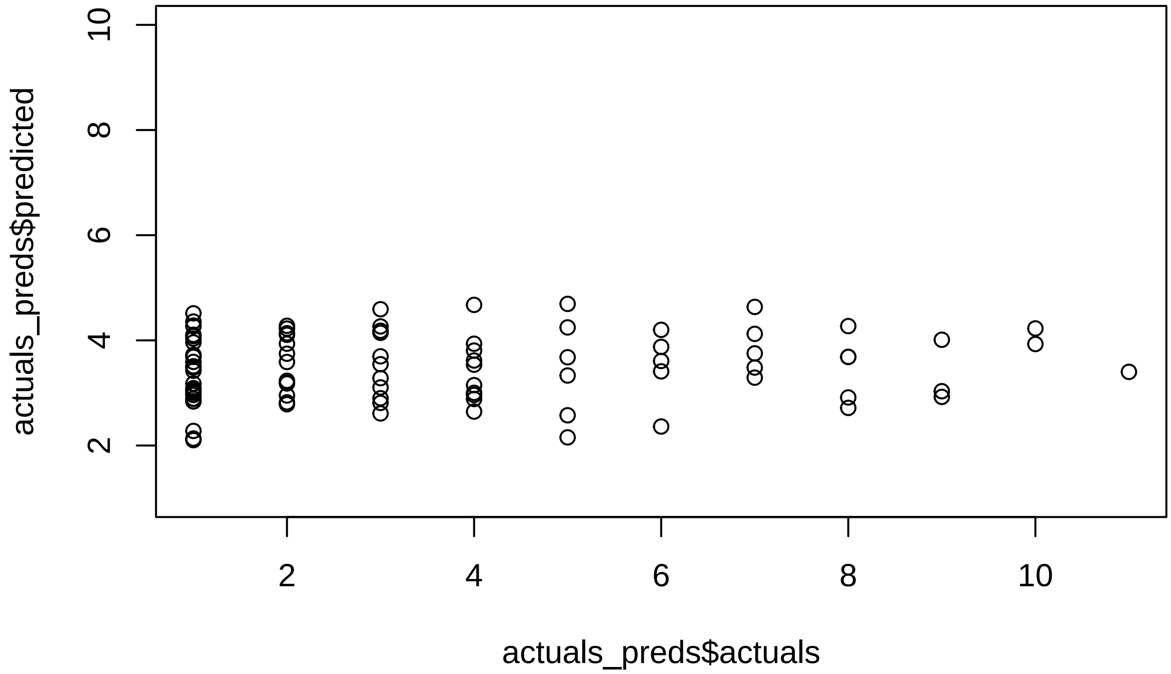
Owing to lack of time in completing this project, I stayed with a linear regression model as a first attempt.

This is a model that fits a straight line through the data.

y = beta_0 + beta_1 * x + error for each data point.

```
##
## Call:
## lm(formula = ElimWeek ~ ., data = trainbachelor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5291 -2.0765 -0.4659  1.5769  7.6185
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.4407607  2.9833854   2.159   0.0320 *
## ...1                 0.0048818  0.0021379   2.283   0.0234 *
## Distance             0.0003612  0.0002775   1.302   0.1945
## ‘Participant’s Age‘ -0.1109266  0.0664323  -1.670   0.0965 .
## ‘Bachelor’s Age‘    -0.0368679  0.0772501  -0.477   0.6337
## id                          NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.632 on 208 degrees of freedom
## Multiple R-squared:  0.04149,    Adjusted R-squared:  0.02305
## F-statistic: 2.251 on 4 and 208 DF,  p-value: 0.06482
```

Let’s look at the plot from this to see where the test elimination week predictions fall:

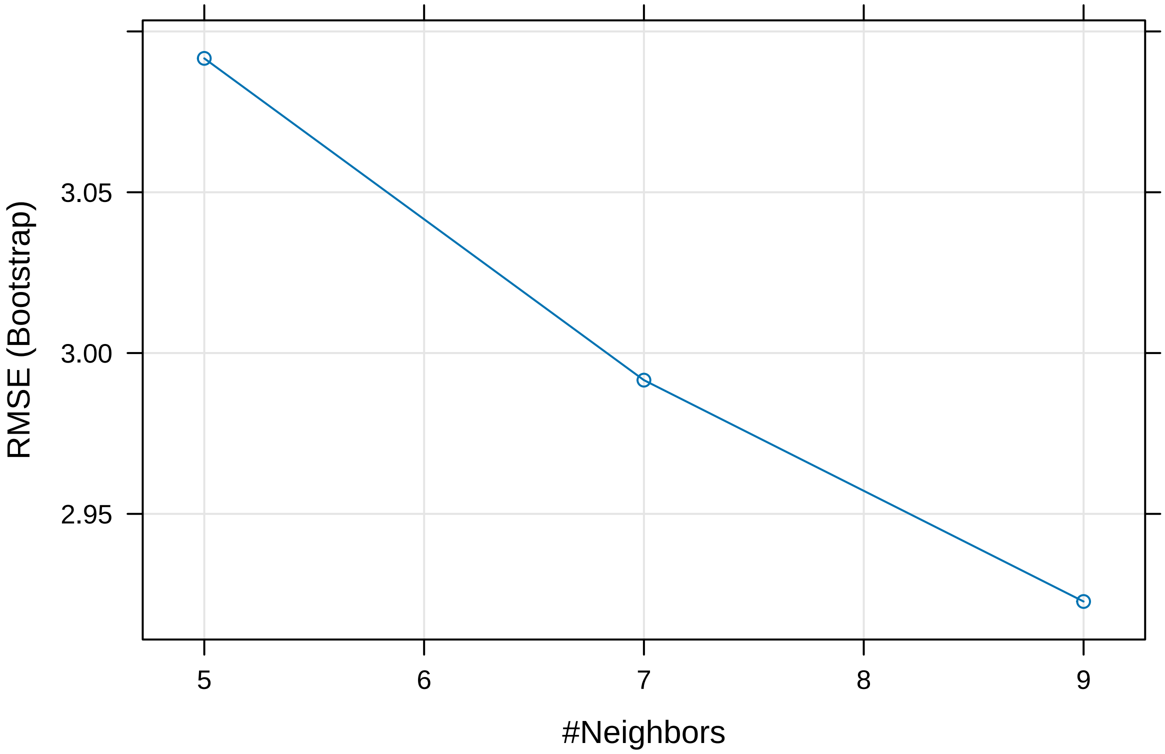


Both basic models indicate the the participant will be eliminated in the third week.

The third method of analysis was knn. The K-nearest neighbors (knn) model, was developed by Fix and Hodges (Hodges, 1951) is a machine learning model that be used for regression analysis. It doesn't make any assumptions about the underlying distribution of the data. It is the simplest algorithm which depends on k neighbors. Owing to its simplicity, the accuracy of using the knn model is not that high and would require a larger dataset than is used in this project

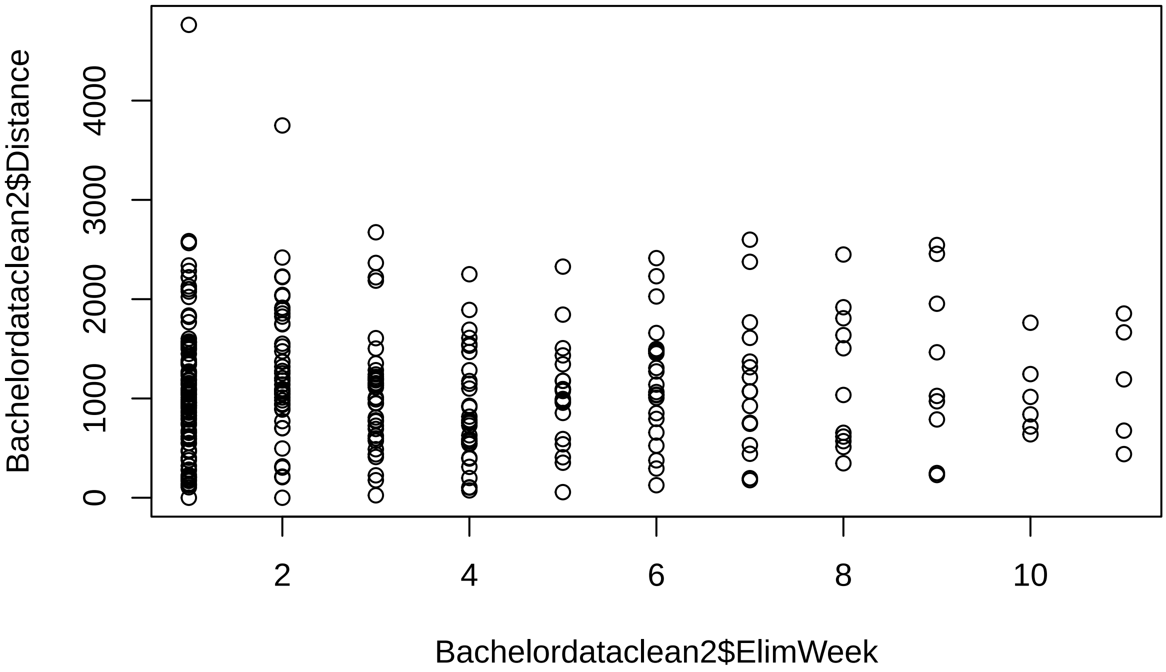
We choose the number of neighbors as k. We add these parameters to the prediction. We want to know if our analysis has improved using the knn model by looking at the rmse from this analysis as well as the confusion matrix.

```
##             Length Class      Mode     
## learn       2      -none-     list     
## k           1      -none-     numeric  
## theDots     0      -none-     list     
## xNames      5      -none-     character
## problemType 1      -none-     character
## tuneValue   1      data.frame list     
## obsLevels   1      -none-     logical  
## param       0      -none-     list     
```

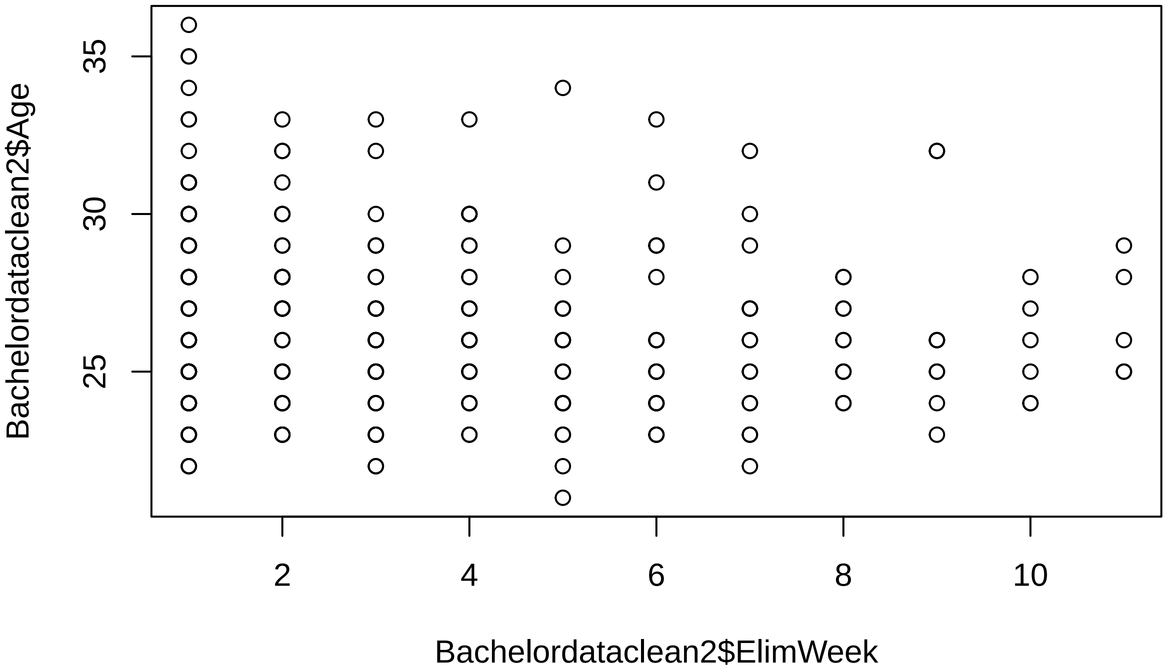


RESULTS

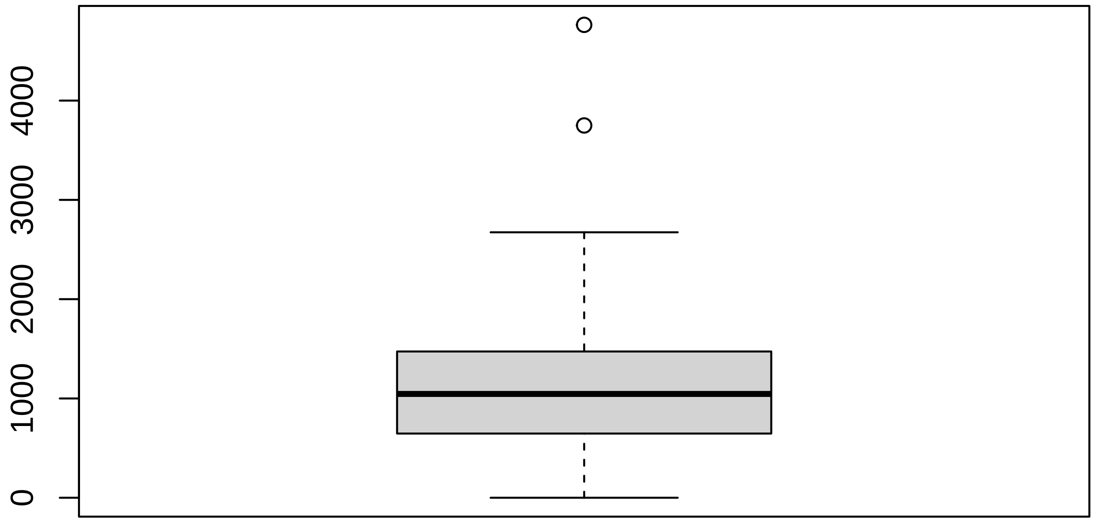
A simple plot of elimination week versus distance shows that if the distance is greater than 2000 miles, the likelihood of a low elimination week is high.



A simple plot of participant's age versus elimination week shows that older participants have lower elimination weeks.



The next plot shows a certain bias towards older participants being eliminated in week 1 to 9, with progressively fewer participants in that age category with each progressive week.



From the boxplot of the distance, we can see two outliers making the distance larger than the other points. A summary gives us the following information: Min. 1st Qu. Median Mean 3rd Qu. Max. 0.0 650.2 1045.6 1104.5 1472.3 4763.1

The mean of the elimination week for the training data was 3.511737. The RMSE from using the mean of the test data is 2.74736. Clearly the mean is not a good predictor of the elimination week as expected.

A summary of age gives the following information:

Min. 1st Qu. Median Mean 3rd Qu. Max. 21.00 24.00 26.00 26.47 28.00 36.00

The minimum age of the participants is 21, the maximum age is 36 and the mean is 26.47.

The first model of linear regression using all the predictors gives us the following result:

Call: lm(formula = ElimWeek ~ ., data = train)

Residuals: Min 1Q Median 3Q Max -3.3454 -2.1047 -0.6835 1.4425 7.7829

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.7192260 3.0167036 2.227 0.027 *. . . 1 0.0032835 0.0021206 1.548 0.123*
Distance 0.0002549 0.0002802 0.910 0.364
`Participant's Age` -0.1515002 0.0661622 -2.290 0.023 `Bachelor's Age` -0.0002499 0.0782499 -0.003 0.997
id NA NA NA NA
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.679 on 208 degrees of freedom Multiple R-squared: 0.04007, Adjusted R-squared: 0.02161 F-statistic: 2.171 on 4 and 208 DF, p-value: 0.07348

If we use only the distance parameter as the predictor, we obtain a slightly different result:

lm(formula = ElimWeek ~ Distance, data = train)

Residuals: Min 1Q Median 3Q Max -3.1854 -2.3779 -0.5411 1.5650 7.6223

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.2441394 0.3721804 8.717 8.53e-16 *** Distance 0.0001976 0.0002828 0.699 0.485
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.712 on 211 degrees of freedom Multiple R-squared: 0.002309, Adjusted R-squared: -0.00242 F-statistic: 0.4883 on 1 and 211 DF, p-value: 0.4855

```
## 
## Call:
## glm(formula = ElimWeek ~ Distance, data = trainbachelor)
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.1728990  0.3598444   8.817 4.43e-16 ***
## Distance    0.0002706  0.0002785   0.971    0.332    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 7.092018)
## 
##     Null deviance: 1503.1  on 212  degrees of freedom
## Residual deviance: 1496.4  on 211  degrees of freedom
## AIC: 1025.7
## 
## Number of Fisher Scoring iterations: 2
```

If you use the generalized linear regression model glm with all the parameters, we get the following summary:

Call: glm(formula = ElimWeek ~ Distance, data = trainbachelor2)

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.478e+00 3.685e-01 9.437 <2e-16 *** Distance -9.976e-05 2.887e-04 -0.346 0.73
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 6.713909)

```
Null deviance: 1417.4  on 212  degrees of freedom
```

Residual deviance: 1416.6 on 211 degrees of freedom AIC: 1014

Number of Fisher Scoring iterations: 2

Giving us an RMSE of 2.87994 indicating a slight improvement in the model with the glm model.

The standard error is larger for this model. The default is the gaussian family or normal distribution. It took 2 iterations to arrive at this result.

Next we used the knn or k nearest neighbors model. This model produced an RMSE of 3.08743.

```
##             Length Class      Mode     
## learn       2      -none-     list     
## k           1      -none-     numeric  
## theDots     0      -none-     list     
## xNames      5      -none-     character
## problemType 1      -none-     character
## tuneValue   1      data.frame list
## obsLevels   1      -none-     logical
## param       0      -none-     list
```

This value is not an improvement over previous values.

CONCLUSIONS

The results in this project indicate that the elimination week of participants in each season of the bachelor is not dependent on the distance between the participant's hometown and the bachelor's hometown.

Age was only a factor in that older participants were eliminated early and had lower elimination weeks.

The least squares, generalized least squares and knn could not confirm any relationship between the distance and the elimination week.

One thing that could be interesting to do if I had more time, would be to add more parameters in the knn analysis and run a confusion matrix to see where there might be some dependence on parameters.

However, there could still be a relationship in other parameters, and this can be investigated. In the knn analysis, we could add a combination of other parameters that are currently available in the current dataset, such as a combination of age, bachelor's age and perhaps a factor or exponential of the distance.

In addition, more data wrangling can obtain more data parameters that were not included in this dataset for this project.

These would include other parameters that are dependent on details of how the show runs each week.

Again, this data is not neatly in one area of the internet and must be gleaned from various other sources, including cleaning the data.

Most importantly, there is still too little data to make predictions and we must wait for more seasons of this show to create much more data on which we might be able to create predictions.

However, in hindsight the data wrangling involved in cleaning this dataset was time-consuming and an already cleaned up dataset on another topic would have provided better results.

REFERENCES

AJ Lee, G. C. (2022). Predicting Winners of the Reality TV Dating Show the Bachelor Using Machine Learning Algorithms. Retrieved from arxiv.org: https://arxiv.org/pdf/2203.16648

Entertainment Betting/ The Bachelor. (n.d.). Retrieved from The Sports Geek: https://www.thesportsgeek.com/entertainment-betting/the-bachelor/

Hodges, E. F. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Randolph Field.

Irizarry, R. (2019). Introduction to data science: Data analysis and prediction algorithms. Chapman and Hall/CRC.

Wiki The Bachelor. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/The_Bachelor_(American_TV_series)