Tony Min
Data Analytics Level 6000
November 30, 2023

Assignment 7

Dataset 1: Absenteeism at Work

1. For Dataset 1, I chose the Absenteeism at Work dataset.
   The first step in the exploratory data analysis is to load the data. In this case the file was separated by semicolon instead of the usual comma. As shown below

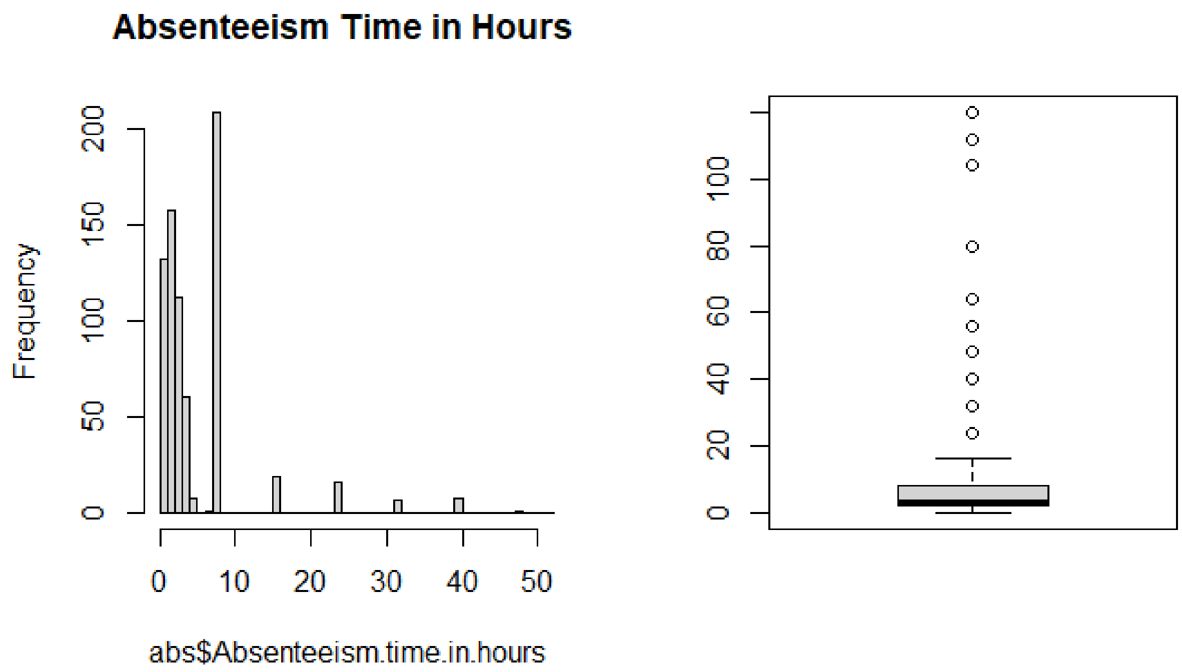
   ```
   abs <- read.csv(file = 'Absenteeism_at_work.csv', sep = ';', header = TRUE)
   ```

   After loading in the data we look at the summaries of all features, using the summary function. After looking at these summary statistics we then will choose some features and look at their distributions and if there seemed to be some outliers.

   The first feature I looked at was the target feature 'Absenteeism.time.in.hours' , the summary statistic is shown below.

   ```
   > summary(abs$Absenteeism.time.in.hours)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     0.000   2.000   3.000   6.924   8.000 120.000
   ```

   As we see the most time absence was 120 hours while the minimum was not absent from work at all with 0 hours. Next we want to see the distributions:



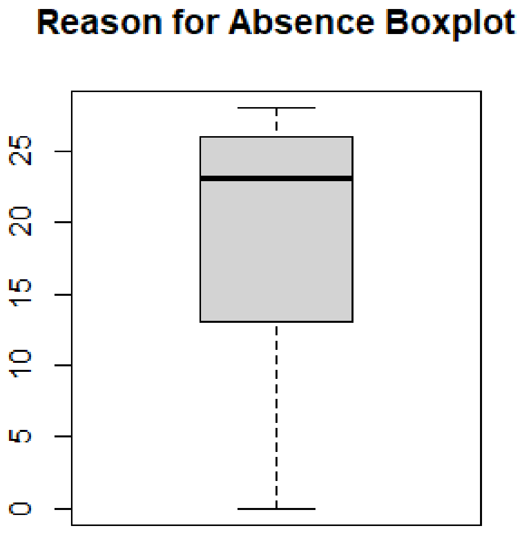
As we see from the boxplot and histogram, the number of hours absent were between 0 to 10 hours. This boxplot is not the full dataset since from the boxplot we see that there were many outliers. I decided to set a limit in the x-axis from 0 to 50 and 100 breaks to get a better view of how many hours. With the mode of hours absent, being around 8.

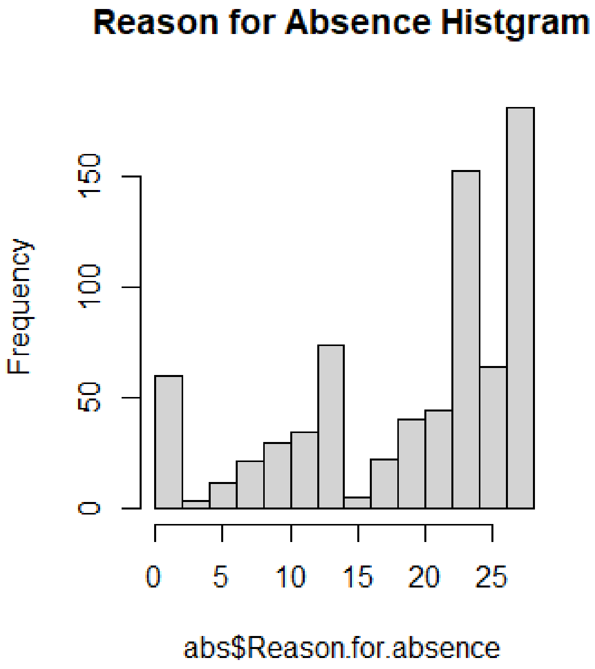
The next feature I took a look at was 'Reason for Absence'. These were classified by the International Code of Diseases which had 21 categories. Which are shown below as key.

I Certain infectious and parasitic diseases
II Neoplasms
III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV Endocrine, nutritional and metabolic diseases
V Mental and behavioural disorders
VI Diseases of the nervous system
VII Diseases of the eye and adnexa
VIII Diseases of the ear and mastoid process
IX Diseases of the circulatory system
X Diseases of the respiratory system
XI Diseases of the digestive system
XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue
XIV Diseases of the genitourinary system
XV Pregnancy, childbirth and the puerperium
XVI Certain conditions originating in the perinatal period
XVII Congenital malformations, deformations and chromosomal abnormalities
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX Injury, poisoning and certain other consequences of external causes
XX External causes of morbidity and mortality
XXI Factors influencing health status and contact with health services.

Now looking at the distributions we can explore the reasons for being absent.

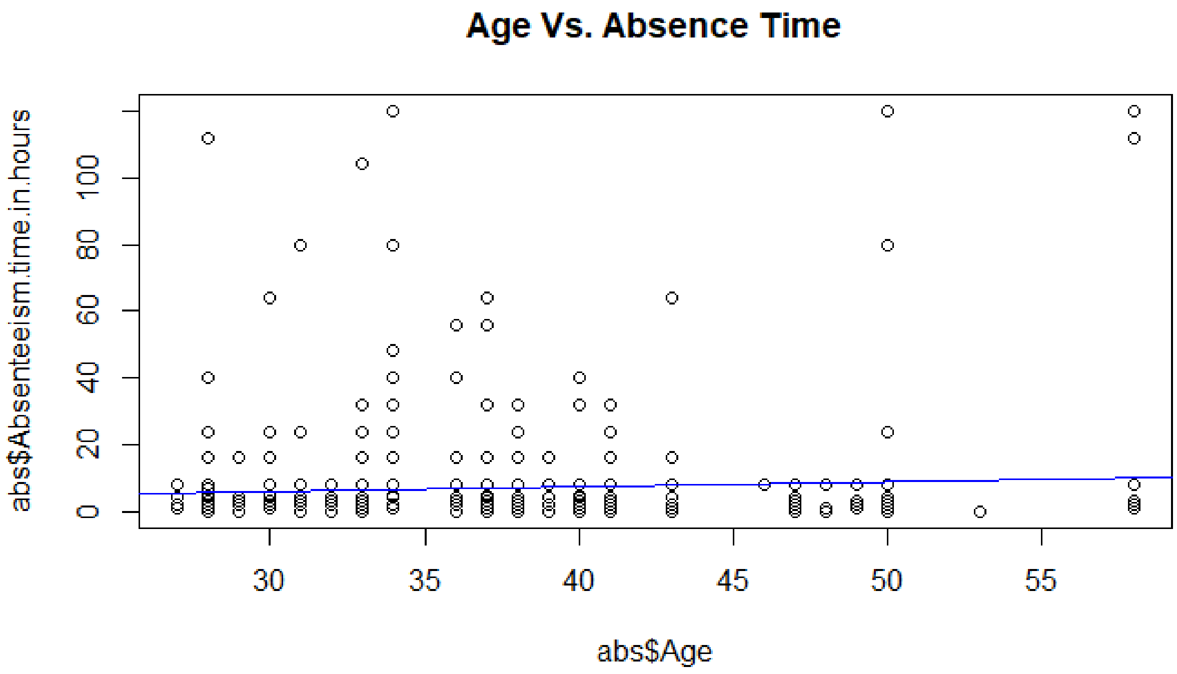




Now looking at these plots we see in the boxplot there are no outliers which makes sense since each individual number relates to a reason to why someone is absent and has no correlation to each other. From the histogram we see most of the reasons were above ‘20’ which correlates to external causes of morbidity and mortality, and factors influencing health status and contact with health services. Then also on the other side of things, the reasons which seemed to be less common were 3 and 15 which correlated to Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism and pregnancy, childbirth and puerperium.

2. Now for the modeling process I decided on Linear Regression, Decision Trees and Random Forest Classification.

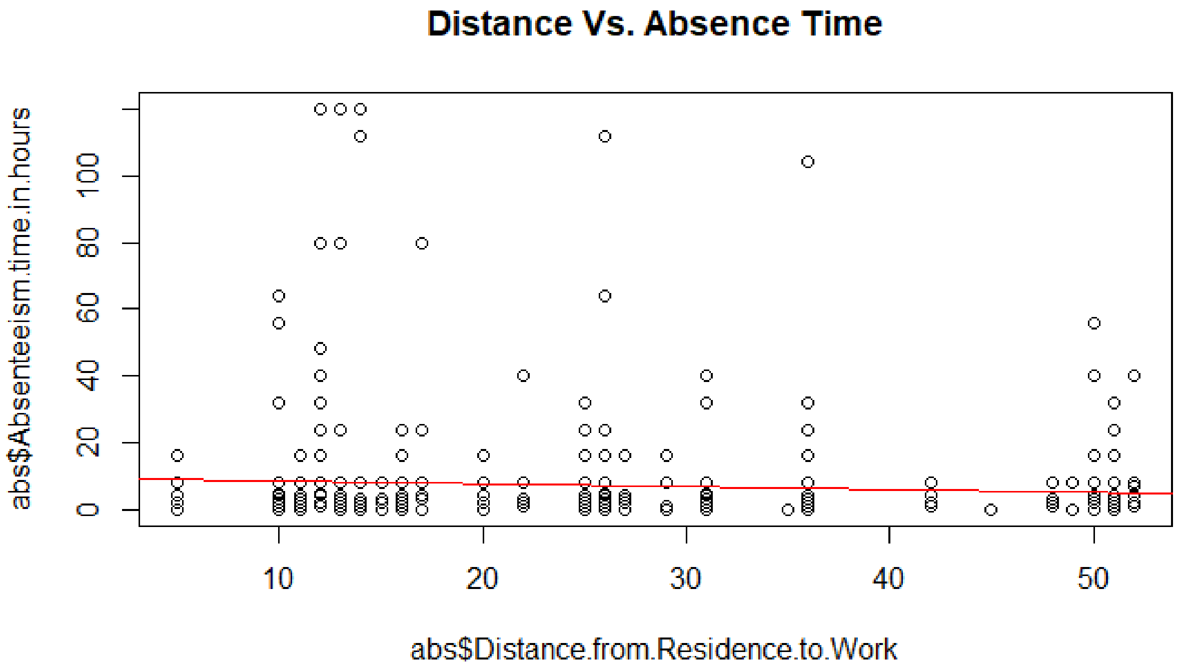
The results of the Linear Regression for Age Vs. Absence Time is shown below:



```
> summary(age)
Call:
lm(formula = abs$Absenteeism.time.in.hours ~ abs$Age)

Residuals:
   Min      1Q  Median      3Q     Max
 -9.164  -5.072  -3.405   1.001 113.407

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99228    2.79797   0.712   0.4767
abs$Age      0.13531    0.07558   1.790   0.0738 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.31 on 738 degrees of freedom
Multiple R-squared:  0.004324,         Adjusted R-squared:  0.002975
F-statistic: 3.205 on 1 and 738 DF,  p-value: 0.07381
```

From the summary of our function we see that we come up with an equation of Absenteeism time in hour = 1.992 + *age* * 0.135, telling us that as we increase in age there the hours absent increase by 0.135. If we look at the p-value calculated 0.073 which means the result is not statistically significant. I thought as the older the person gets the more absent they may be but from this model it shows that there is no correlation between age of the worker and the number of hours they are absent from work. Also in the dot plot we see that the line seems to almost be a straight line with a slight incline.

The next linear regression model I did was Distance from work Vs. Absence time.

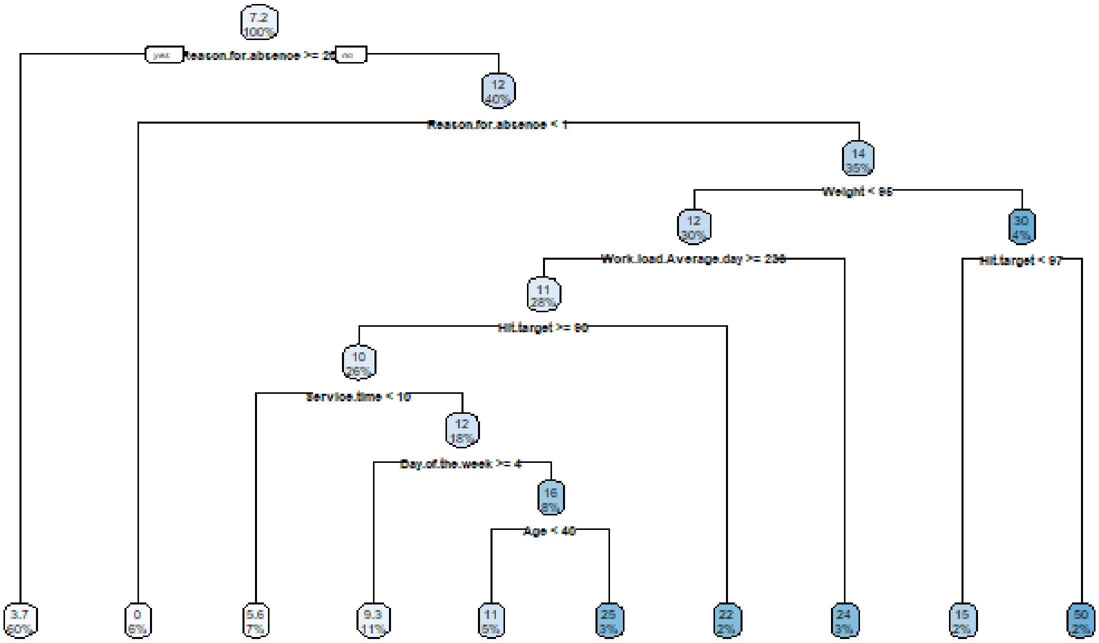


```
> summary(dist)
Call:
lm(formula = abs$Absenteeism.time.in.hours ~
abs$Distance.from.Residence.to.Work)

Residuals:
   Min      1Q  Median      3Q     Max
 -8.880  -5.245  -3.228   0.708 111.835

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          9.27688    1.09159   8.498   <2e-16 ***
abs$Distance.from.Residence.to.Work -0.07939    0.03295  -2.410   0.0162 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 13.29 on 738 degrees of freedom
Multiple R-squared:  0.007808,	Adjusted R-squared:  0.006464
F-statistic: 5.808 on 1 and 738 DF,  p-value: 0.0162
```

From the summary we get a linear equation of Absenteeism time in hours = 9.277 + *Distance from work* * (-0.079). Meaning an increase in distance from work will result in a decrease in absence from work by 0.079 hours. From this equation and the graph we see that there seemed to be a small negative correlation with distance from work and the number of absentee hours. Also looking at the p-value of 0.016, we can see that it is statistically significant, therefore there is a direct negative correlation between the two variables.

The next model was a Decision Tree which is shown below



Looking at this decision tree we see that at the start we have the root which is the reason for absence based on the condition if it is greater or equal to 12. Then further down the tree some other nodes that determined the absentee hours were weight, workload, hit target, service time, day of week and age. With this visual we see that these are some of the most important factors in determining the absentee hours.

The final model was a Random Forest Regression.

```
> rf

Call:
 randomForest(formula = Absenteeism.time.in.hours ~ ., data = TrainSet,
importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 139.5195
                    % Var explained: 8.02
```

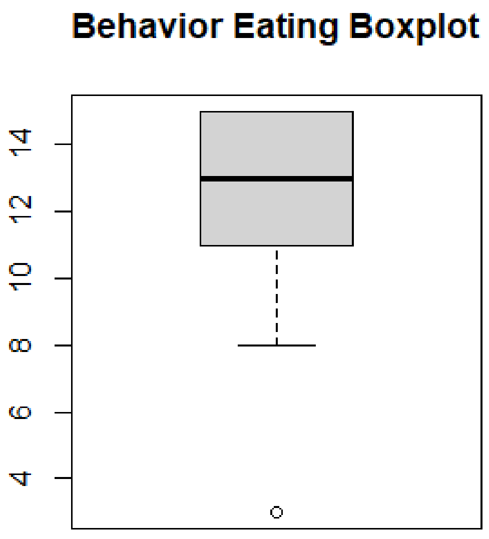
From the results of running and creating a random forest regression model we can see how we built a forest with 500 trees and at each split the algorithm used 6 randomly selected variables. The percent of variance explained for ‘Absenteeism time in hours’ is only 8.02%, which isn’t a lot. But next we use importance(rf) to get the important variables. We find that some of the important variables with a positive Increase in Mean Squared Error were Reason for Absence, Disciplinary Failure and Average Workload. Vice versa, some features that had a negative Increase in Mean Squared Error were Social Smoker, Pet and Age.

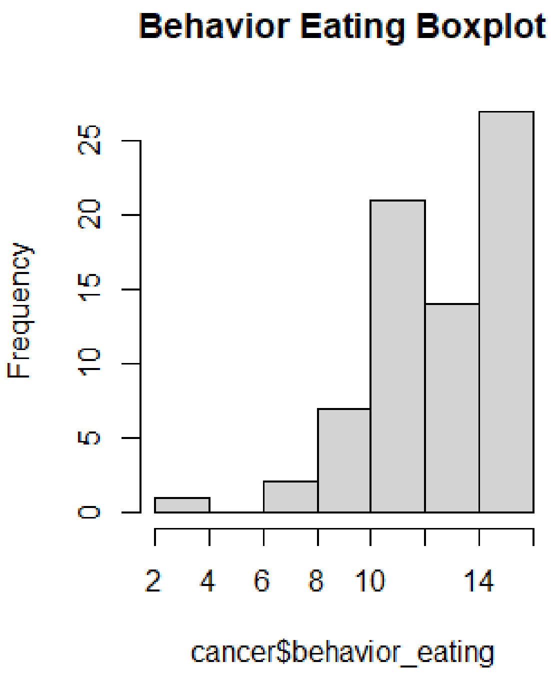
3. Firstly looking at the linear regression it is easy to see the significance of each variable to the target feature ‘absenteeism time in hours’, the two features we chose we saw how age was not statistically significant in predicting the target feature. This was then further proven in the random forest regression as Age was a feature in a negative Increase in Mean Squared Error. In the next linear regression model we made, comparing distance from work to absenteeism time in hours we see there was a negative correlation between them but was statistically significant which was a significant feature in the random forest classifier, having a positive Increase in Mean Squared Error. I find it interesting how Age was not a significant feature since before this I thought the older the worker the more time absent then would be. On the other hand, the negative correlation between distance from work and time absent was also interesting, in my mind I thought if someone worked farther away they may tend to miss more work. In the future I think it would be helpful to perform a multivariate regression on variables that were statistically significant and then be able to get more insight on how those significant variables relate to each other. Now onto the decision tree it was helpful to see the distribution of data points in relation to all the different variables, but visually it is slightly confusing. They didn’t really explain certain variables such as the reason for absence, they gave a general description instead of a specific reason. Another confusing feature from the decision tree was hit target since the website did not explain this variable at all. Now from the Random Forest Regression I found it interesting when using all the features in that it only explained 8.02% of the

variance in ‘absenteeism time in hours’. While it was surprising I think if you run it again without the features that weren’t statistically important we may be able to cover a higher percentage of variance within the model. Although I don’t know how much more variance it could cover since it seemed that only 7 variables had a negative Increase in Mean Squared Error.

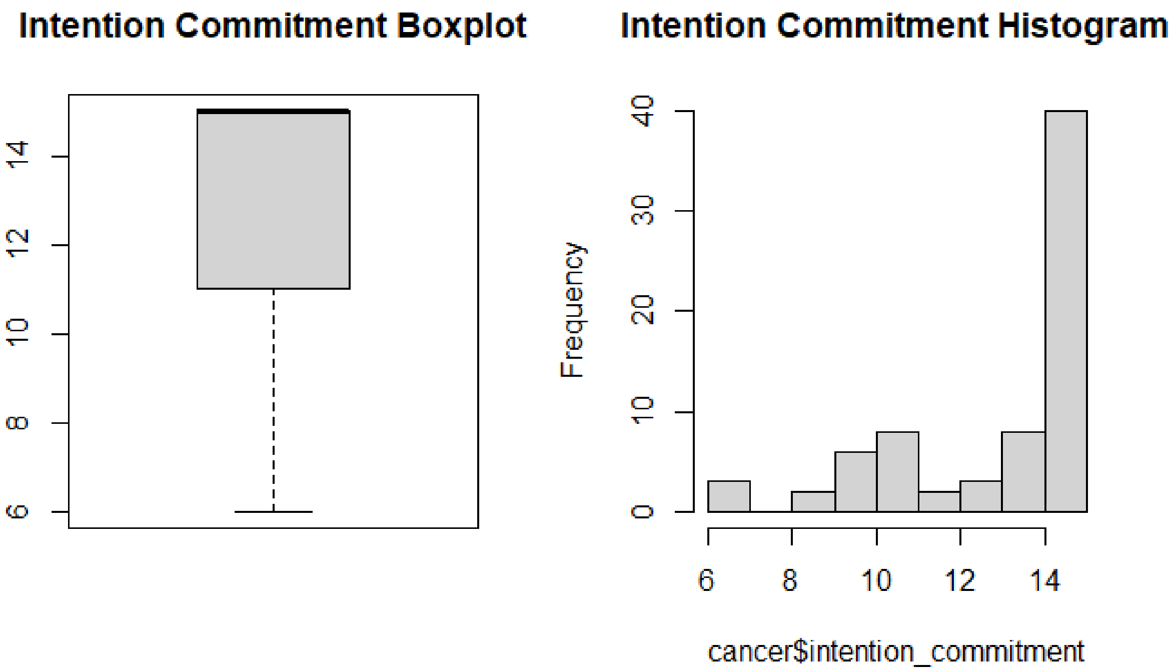
Dataset 2: Cervical Cancer Behavior Risk

1. For the second dataset I chose the Cervical cancer Behavior Risk dataset. For the exploratory data analysis we load in the data, which this time was formatted correctly so simply running the *read.csv()* function. Once the data was loaded we ran the *summary()* function on the whole dataset to get a basic idea of the data. Along with looking at the summary of the dataset I used the website from where the dataset was downloaded and read about each of the variables in the columns. We then randomly choose a couple of variables to learn about their distributions.





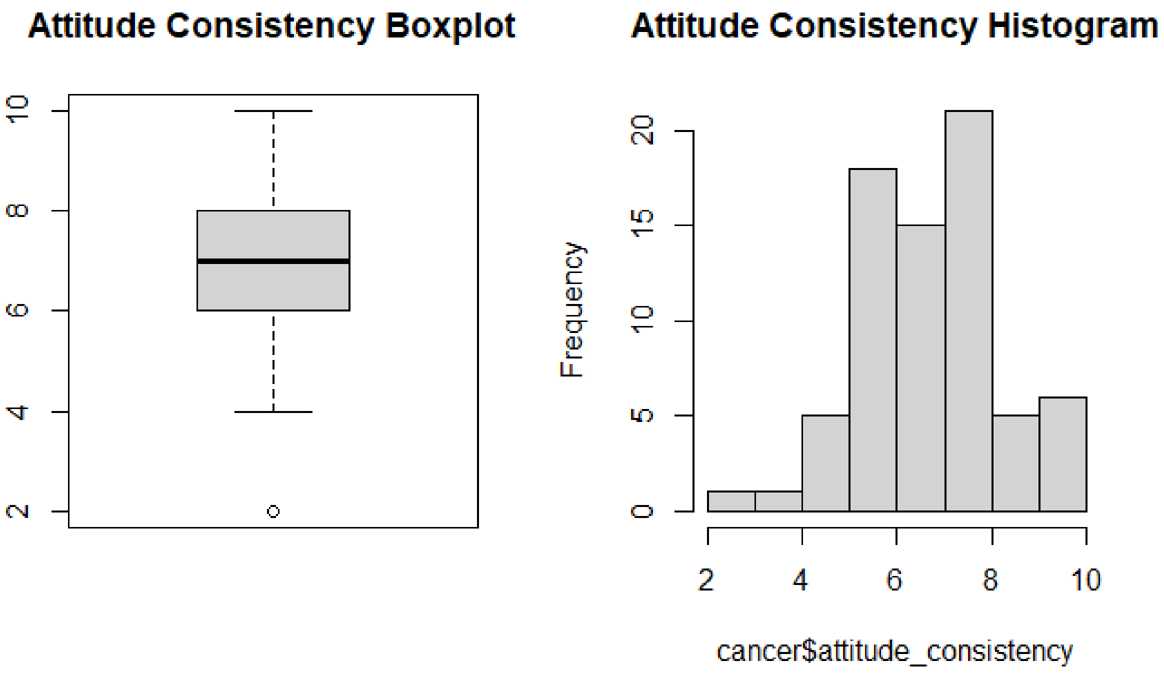
Looking at the boxplot and histogram of behavior eating it seems like there is an outlier below 4 in the boxplot. On the histogram we see that there is a high concentration of values of 10+ in the dataset for behavior_eating. While we don’t know what these values actually mean, in my opinion this is about how much they eat in a certain period of time, could be meals in a week.



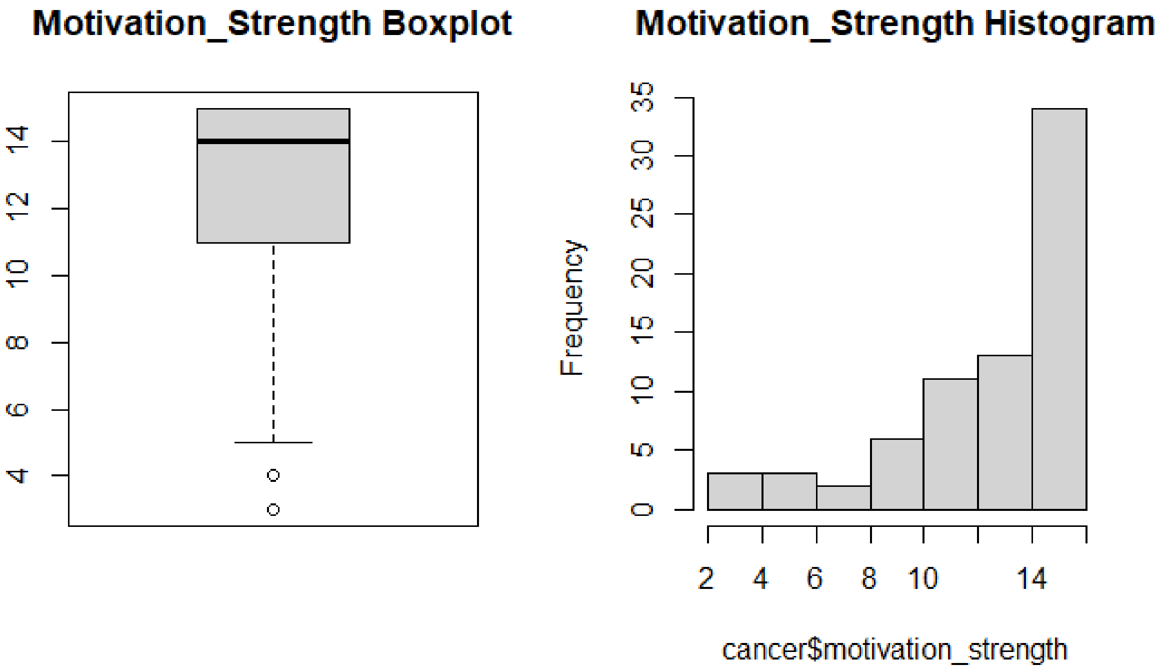
Intention Commitment Boxplot

Intention Commitment Histogram

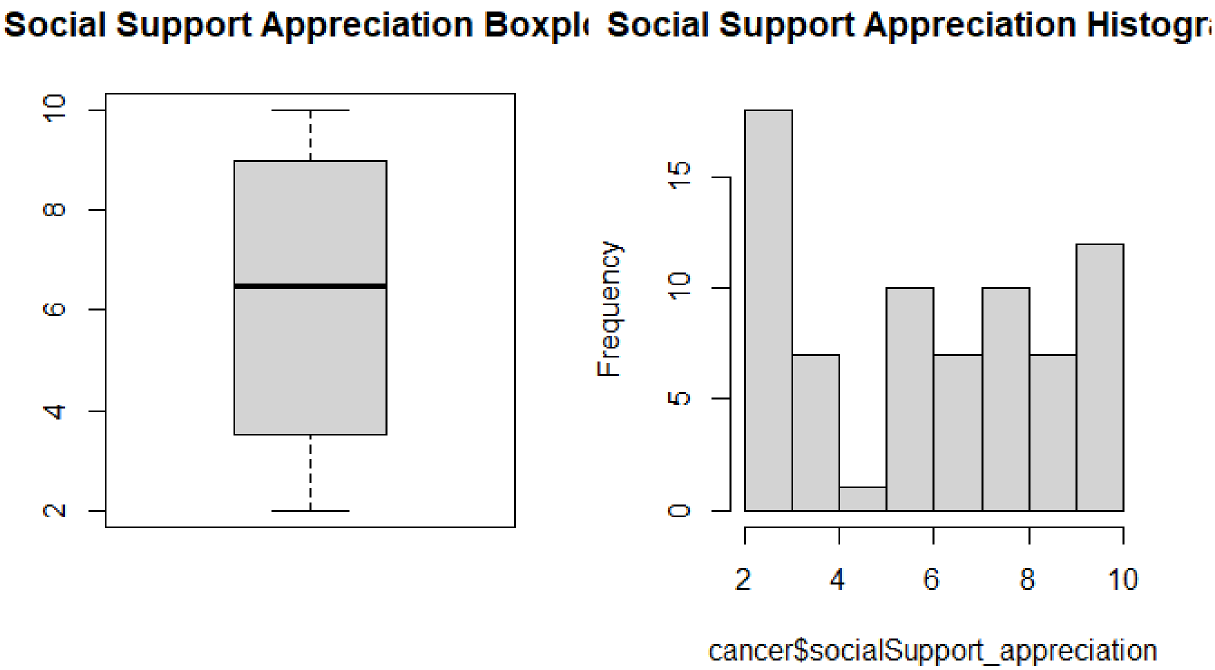
Now looking at the intention_commitment variables we can see there are no outliers and a range from 6 to 15. While again the website does not specify what these values mean I suspect that these correlate to the commitment levels of some sort. Based on this theory it seems that many individuals have a high commitment level as shown in the histogram and we can gather that from the boxplot since the mean and max are both at 15.



For this variable, attitude_consistency, we see that this distribution is more normal than the other two variables we have looked at. Seems like the cluster of individuals are around 5-8 range. Again we don't know what these values correlate to from the website.



There are 2 outliers for the motivation_strength variable. From the histogram we see how the data is skewed towards the right, the 15 being the most frequent.


Social Support Appreciation Boxpl

Social Support Appreciation Histogr

For socialSupport_aprreciation there seemed to be no outliers. Looking at the histogram we see that the highest frequency is 2 with the lowest being 4, all other values seemed to have similar frequencies.