Final Project

Hagay Ringel

2022-12-15

Background: At this project, I will analyze worldwide alcohol consumption and its impact on happiness by using two data sets from Kaggle. The first data set engage with the level of happiness as a result of alcohol consumption, specifically beers, spirits and wine, We can also find data about the countries and their HDI (Human Development Index). The second data set includes data about the worldwide alcohol consumption by gender.

This project has two parts: In the first part, I will visual and produce linear regression in order to examine whether beer has an impact on the happiness level and showing whether developed countries consume more beer than poor countries. In the second part of the project I will address the question: Can we predict the sex by alcohol consumption level?

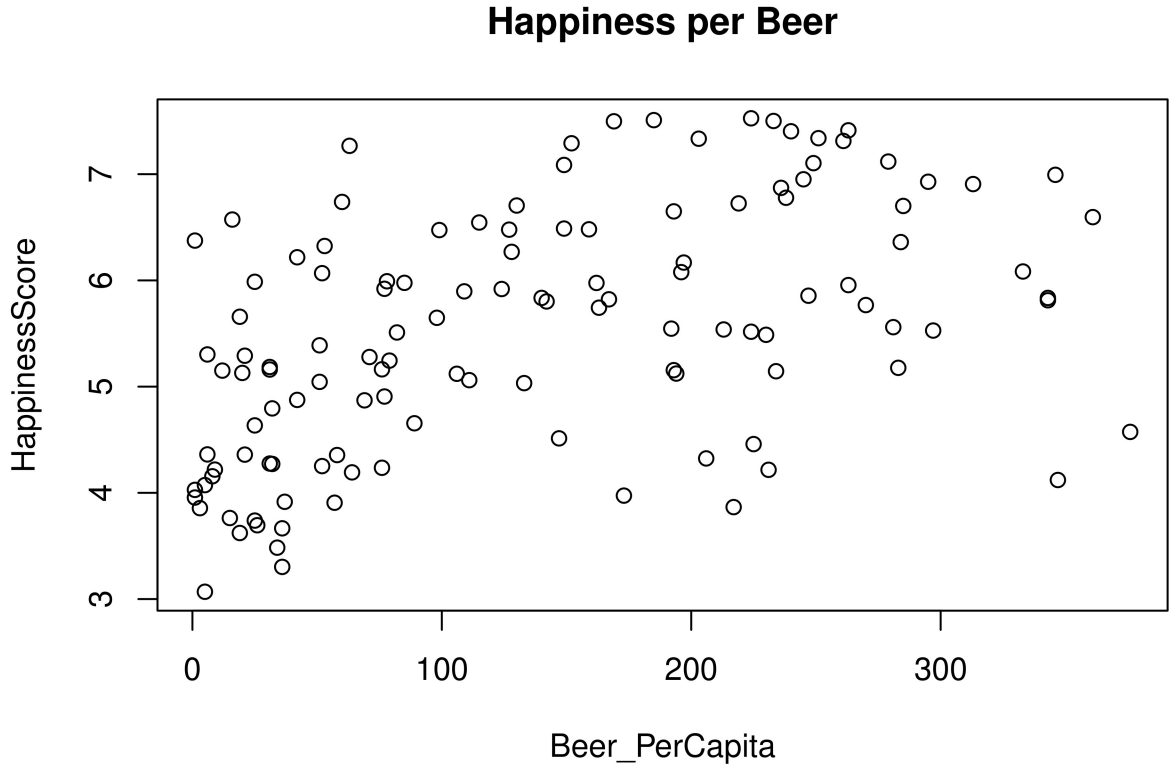
In this project, I will use techniques I learned in class, such as linear regression, visualization, data cleansing, merging data sets, functions of confusion matrix and using the algorithms KNN and Decision Tree for the ML portion. All of this in order to try to predict gender based on alcohol consumption: My assumption is that males drink more than females, therefore, we will be able to assume that above X quantity of alcohol it is more likely to be a male.

```
happiness = read.csv("C:/Program Files/R/Happiness.csv")
```

Cleaning the data

```
happiness = na.omit(happiness)
```

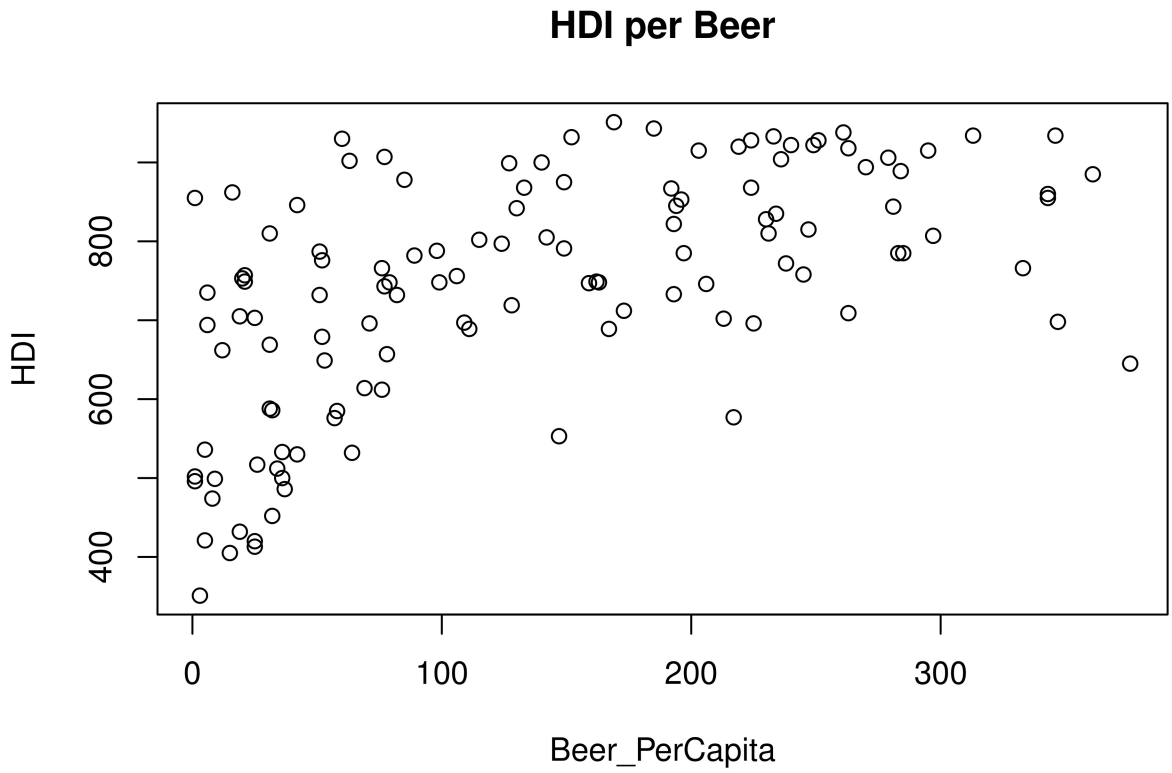
Showing by graph and linear regression whether alcohol have an effect on the level of happiness




```
Call:
lm(formula = HappinessScore ~ Beer_PerCapita, data = happiness)

Residuals:
     Min       1Q   Median       3Q      Max
-2.53624 -0.73772  0.04062  0.79423  2.14535

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.7810091  0.1502750  31.815  < 2e-16 ***
Beer_PerCapita 0.0054070  0.0008702   6.213 7.75e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.003 on 120 degrees of freedom
Multiple R-squared:  0.2434,    Adjusted R-squared:  0.2371
F-statistic: 38.61 on 1 and 120 DF,  p-value: 7.752e-09
```

Analysis: Since we have a high P- value (7.75e-09) and low R-squared (0.2371), we can assume that model doesn't explain much of variation of the data and it is not significant. The graph is also a tool to see that when the dota are too scattered.

Showing by graph and linear regression whether developed countries consume more beer




```
Call:
lm(formula = HDI ~ Beer_PerCapita, data = happiness)

Residuals:
    Min      1Q  Median      3Q     Max
-296.44  -77.93   15.38   81.58  254.37

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    625.1619    18.2223  34.308  < 2e-16 ***
Beer_PerCapita   0.8412     0.1055   7.971 1.03e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 121.7 on 120 degrees of freedom
Multiple R-squared:  0.3462,    Adjusted R-squared:  0.3408
F-statistic: 63.54 on 1 and 120 DF,  p-value: 1.027e-12
```

Analysis: Since we have a high P- value (1.03e-12) and low R-squared (0.3408), we can assume that model doesn't explain much of variation of the data and it is not significant.

```
alco_who = read.csv("C:/Program Files/R/WHOAlcohol.csv")
```

Cleaning the data

```
alco_who = na.omit(alco_who)
```

Since the first data set is from 2016, let's filter the second data set according to the same year- 2016

```
alco_who2 = alco_who %>% filter(alco_who$Year == 2016)
```

Changing some columns names to more readable names

```
names(alco_who2)[7] ="alcohol_normal"
names(alco_who2)[8] ="alcohol_low"
names(alco_who2)[9] ="alcohol_high"
names(alco_who2)[10] ="alcohol_string"
```

Merging the happiness data set with alco_who2 data set

```
merged_table = merge(happiness, alco_who2)
```

Cleaning the new merged data set

```
merged_table = na.omit(merged_table)
```

Converting the gender to numeric values

```
merged_table$Sex <- c('Female' = 0, 'Male' = 1, 'Both sexes' = 2)
```

Since the value "Both sexes" is unnecessary for our prediction and we need to keep it binary (two values only), we better to remove it

```
merged_table = merged_table %>% filter(merged_table$Sex < 2)
```

Setting functions for sensitivity, specificity, accuracy and precision

```
sensitivity = function(cm) {
  return(cm[2,2]/(cm[2,2]+cm[2,1]))
}
```

```
specificity = function(cm) {
  return(cm[1,1]/(cm[1,2]+cm[1,1]))
}
```

```
accuracy = function(cm) {
  return((cm[1,1]+cm[2,2])/(cm[1,1]+cm[1,2]+cm[2,1]+cm[2,2]))
}
```

```
precision = function(cm) {
  return(cm[2,2]/(cm[2,2]+cm[1,2]))
}
```

Setting the variable we want to predict

```
merged_table$Sex = factor(merged_table$Sex,
                      levels=c(0,1),
                      labels=c("Female","Male"))
```

Normalizing the relevant columns

```
merged_table$Beer_PerCapita = scale(merged_table$Beer_PerCapita)
merged_table$Spirit_PerCapita = scale(merged_table$Spirit_PerCapita)
merged_table$Wine_PerCapita = scale(merged_table$Wine_PerCapita)
```

Creating a training data set corresponding to 70% of the available data

```
ind = sample(2, nrow(merged_table), replace=TRUE, prob=c(0.7, 0.3))
```

```
merged_training = merged_table[ind==1, 7:9]
merged_test = merged_table[ind==2, 7:9]
merged_trainLabels = merged_table[ind==1, 14]
merged_testLabels = merged_table[ind==2, 14]
```

Performing a KNN prediction of Sex as a function of Beer_PerCapita, Spirit_PerCapita and Wine_PerCapita

```
prediction = knn(train = merged_training,
                 test = merged_test,
                 cl = merged_trainLabels,
                 k = 3)
```

Producing confusion matrix

```
(confusionMatrix = table(Actual_Value = merged_testLabels,
                         Predicted_Value = prediction))
```

```
            Predicted_Value
Actual_Value Female Male
      Female     12   26
      Male       25   12
```

Calculating sensitivity, specificity, accuracy and precision

```
sensitivity(confusionMatrix)
```

```
[1] 0.3243243
```

```
specificity(confusionMatrix)
```

```
[1] 0.3157895
```

```
accuracy(confusionMatrix)
```

```
[1] 0.32
```

```
precision(confusionMatrix)
```

```
[1] 0.3157895
```

Analysis: According to the KNN algorithm, the number of the FN is high as well as the accuracy, therefore, we can't predict the sex by alcohol consumption and we can rely on the result since the accuracy percentage is low.

Performing a Decision Tree prediction of Sex as a function of Beer_PerCapita, Spirit_PerCapita and Wine_PerCapita

```
trainingWithLabel = merged_training
trainingWithLabel$Sex = merged_trainLabels
```

```
model = rpart(Sex ~ Beer_PerCapita +
                Spirit_PerCapita +
                Wine_PerCapita,
              data=trainingWithLabel,
              control=rpart.control(maxdepth=3),
              method='class')
```

```
prediction = predict(model, merged_test, type='class')
```

Producing the confusion matrix

```
(confusionMatrix = table(Actual_Value = merged_testLabels,
                         Predicted_Value = prediction))
```

```
            Predicted_Value
Actual_Value Female Male
      Female     13   25
      Male       24   13
```

Calculating sensitivity, specificity, accuracy and precision

```
sensitivity(confusionMatrix)
```

```
[1] 0.3513514
```

```
specificity(confusionMatrix)
```

```
[1] 0.3421053
```

```
accuracy(confusionMatrix)
```

```
[1] 0.3466667
```

```
precision(confusionMatrix)
```

```
[1] 0.3421053
```

Analysis: According to the Decision Tree algorithm, the number of the FN is high (16) as well as the accuracy (0.29), therefore, we can't predict the sex by alcohol consumption. This result means a high probability that

something is missed and the accuracy rate is low which means a low percentage of the correctness of this prediction- which mean we can rely on this result.

Conclusion: According to the results of this project, we can say that there is no relationship between the happiness level to drinking beer, as well as no relationship between the development level of the country to beer consumption. As for the second part of this project, we can't predict the gender by the amount of alcohol the person drinks, in other words: there is no significant difference in alcohol consumption between male and female. A suggestion for a further research: Using a larger data set may lead to a different results, however, according to this research my hypothesis at the beginning of the paper has disproved.