DA5030.P1.Sabbisetty.Rmd

2023-10-10

Problem 1

1 Predicting Life Expectancy

```
## ’data.frame':    2938 obs. of  20 variables:
##  $ Country              : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                 : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status               : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ LifeExpectancy       : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ AdultMortality       : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ NumInfantDeaths      : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol              : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ PercentageExpenditure: num  71.3 73.5 73.2 78.2 7.1 ...
##  $ HepB                 : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles              : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                  : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ Under5Deaths         : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ TotalExpenditure     : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria           : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV                  : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                  : num  584.3 612.7 631.7 670 63.5 ...
##  $ thinness1.19y        : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness5.9y         : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Schooling            : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```
##        Country Year     Status LifeExpectancy AdultMortality NumInfantDeaths
## 1  Afghanistan 2015 Developing           65.0            263              62
## 2  Afghanistan 2014 Developing           59.9            271              64
## 3  Afghanistan 2013 Developing           59.9            268              66
## 4  Afghanistan 2012 Developing           59.5            272              69
## 5  Afghanistan 2011 Developing           59.2            275              71
## 6  Afghanistan 2010 Developing           58.8            279              74
## 7  Afghanistan 2009 Developing           58.6            281              77
## 8  Afghanistan 2008 Developing           58.1            287              80
## 9  Afghanistan 2007 Developing           57.5            295              82
## 10 Afghanistan 2006 Developing           57.3            295              84
##    Alcohol PercentageExpenditure HepB Measles  BMI Under5Deaths Polio
## 1     0.01             71.279624   65    1154 19.1           83     6
## 2     0.01             73.523582   62     492 18.6           86    58
## 3     0.01             73.219243   64     430 18.1           89    62
## 4     0.01             78.184215   67    2787 17.6           93    67
## 5     0.01              7.097109   68    3013 17.2           97    68
## 6     0.01             79.679367   66    1989 16.7          102    66
## 7     0.01             56.762217   63    2861 16.2          106    63
## 8     0.03             25.873925   64    1599 15.7          110    64
## 9     0.02             10.910156   63    1141 15.2          113    63
## 10    0.03             17.171518   64    1990 14.7          116    58
##    TotalExpenditure Diphtheria HIV       GDP thinness1.19y thinness5.9y
## 1              8.16         65 0.1 584.25921          17.2         17.3
## 2              8.18         62 0.1 612.69651          17.5         17.5
## 3              8.13         64 0.1 631.74498          17.7         17.7
## 4              8.52         67 0.1 669.95900          17.9         18.0
## 5              7.87         68 0.1  63.53723          18.2         18.2
## 6              9.20         66 0.1 553.32894          18.4         18.4
## 7              9.42         63 0.1 445.89330          18.6         18.7
## 8              8.33         64 0.1 373.36112          18.8         18.9
## 9              6.73         63 0.1 369.83580          19.0         19.1
## 10             7.43         58 0.1 272.56377          19.2         19.3
##    Schooling
## 1       10.1
## 2       10.0
## 3        9.9
## 4        9.8
## 5        9.5
## 6        9.2
## 7        8.9
## 8        8.7
## 9        8.4
## 10       8.1
```

```
##    Country               Year         Status          LifeExpectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  AdultMortality  NumInfantDeaths     Alcohol       PercentageExpenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##       HepB          Measles              BMI        Under5Deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.05
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34      NA's   :1
##      Polio       TotalExpenditure   Diphtheria         HIV
##  Min.   : 3.00   Min.   : 0.370   Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260   1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755   Median :93.00   Median : 0.100
##  Mean   :82.54   Mean   : 5.938   Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492   3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600   Max.   :99.00   Max.   :50.600
##  NA's   :21      NA's   :226      NA's   :19
##       GDP            thinness1.19y    thinness5.9y     Schooling
##  Min.   :     1.68   Min.   : 0.10   Min.   : 0.10   Min.   : 0.00
##  1st Qu.:   463.94   1st Qu.: 1.60   1st Qu.: 1.50   1st Qu.:10.10
##  Median :  1766.95   Median : 3.30   Median : 3.30   Median :12.30
##  Mean   :  7483.16   Mean   : 4.84   Mean   : 4.87   Mean   :11.99
##  3rd Qu.:  5910.81   3rd Qu.: 7.20   3rd Qu.: 7.20   3rd Qu.:14.30
##  Max.   :119172.74   Max.   :27.70   Max.   :28.60   Max.   :20.70
##  NA's   :448         NA's   :34      NA's   :34      NA's   :163
```

1.1 / Analysis of Data Distribution

```
## [1] 73
```

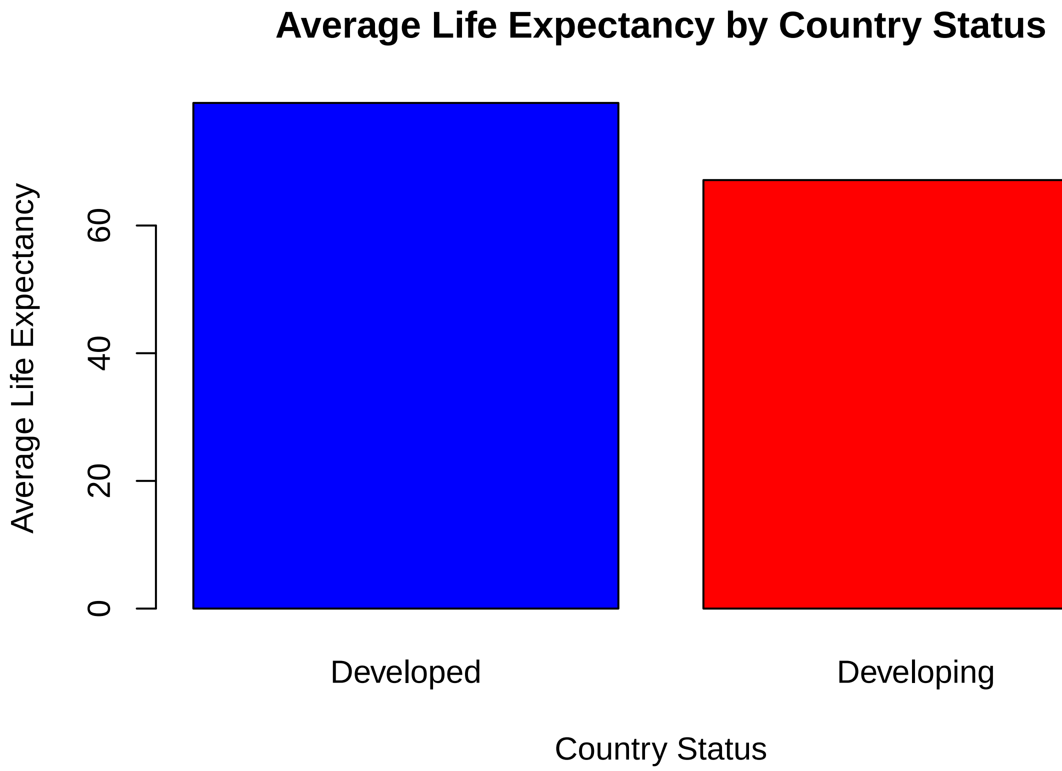
```
## [1] 69.22493
```

```
## [1] 72.1
```

```
## [1] 9.523867
```

#Developed Countries: The average life expectancy for developed countries is approximately 79.197 years. #Developing Countries: In contrast, the average life expectancy for developing countries is notably lower, at around 67.11 years. #This analysis reveals a significant difference in average life expectancy between these two categories of countries. Developed countries tend to have a significantly higher average life expectancy compared to developing countries. This difference could be attributed to various factors, such as access to healthcare, socioeconomic conditions, and public health measures. #The bar chart would visually represent this contrast, making it easier to grasp the disparities in life expectancy between “Developed” and “Developing” countries.

```
##       Status LifeExpectancy
## 1  Developed       79.19785
## 2 Developing       67.11147
```



#Question 3 - (3 pts) Adding to the above question, determine if the difference in mean life expectency between the two types of countries is statistically significant. Add appropriate writing to your report and use embedded code to ensure that the report is updated if the data changes.

```
##
##  Welch Two Sample t-test
##
## data:  LifeExp.df$LifeExpectancy by LifeExp.df$Status
## t = 47.868, df = 1807, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Developed and group Developing is not
## 95 percent confidence interval:
##  11.59118 12.58159
## sample estimates:
##  mean in group Developed mean in group Developing
##                 79.19785                 67.11147
```

The t-statistic for this analysis is approximately 47.868, and the degrees of freedom are 1807. The p-value is remarkably small, well below the commonly used significance level of 0.05 (p-value < 2.2e-16).

In the context of our hypotheses: - Null Hypothesis (H0): There is no significant distinction in the average life expectancy between Developed and Developing countries. - Alternative Hypothesis (H1): There is a significant difference in average life expectancy between Developed and Developing countries.

Given the extremely small p-value, we have strong grounds to reject the null hypothesis and favor the alternative hypothesis. This provides robust statistical evidence that there is indeed a significant variance in average life expectancy between Developed and Developing countries.

Moreover, the 95 percent confidence interval for the difference in means (ranging from 11.59118 to 12.58159) indicates that we can be highly confident the true difference in average life expectancy between these two categories falls within this range. Additionally, the means for the two groups (79.19785 for Developed and 67.11147 for Developing) show a substantial discrepancy.

In summary, the data strongly supports the assertion that there is a meaningful difference in average life expectancy between Developed and Developing countries, with Developed countries exhibiting a notably higher average life expectancy.

Question 4 - (5 pts) Test the normality of the column "life expectancy" by performing either a Shapiro-Wilk (tutorial Links to an external site.) or Kolmogorov-Smirnov Links to an external site. test. Describe what you found in markdown. Do not echo the code, just include the result in markdown. Be sure to add your analysis of the p-value.

```
##
##  Shapiro-Wilk normality test
##
## data:  LifeExp.df$LifeExpectancy
## W = 0.95605, p-value < 2.2e-16
```

The object NormalityTest.LifeExp returned from the function shapiro.test() contains the value of *W in 0.95* and the *p-value in < 0.05*.

The data exhibits non-normal distribution based on the Shapiro-Wilk normality test (*W in 0.95* , *p < 0.05*).

1.2 / Identification of Outliers.

Which are your outliers for each column? What would you do? Summarize the results of your analysis and any potential strategies in your notebook's markdown. Explain how you identified the outliers and how many you found. What were the max and min for life expectancy? What about the standard deviation? What is the median? Would it make sense to calculate a trimmed mean? How would you use your analysis of outliers to determine the percentage to trim? Explain all of that and the results.

The max and min for life expectancy are 89 and 36.3 respectively.

The median Life Expectency is 72.1 with a standard deviation of 9.52. The average Life Expectency was 69.22.

Z score values are used to identify outliers. Z score value is obtained by (column mean - each value in column)/ standard deviation. For any column, if the z score value is above 3 standard deviations, the value is considered as an outlier.

Under Data Prepration I would normalize the data with z score standardization method which is essential to maintain low variance especially for distance based algorithms like knn.

Trimmed mean can be performed but might result in loss of data, thus instead I chose to identify outliers, convert to NA and the mean impute missing values.

Below are the outliers for each column:

a) For column Life.expectancy total number of outliers are 2.

b) For column AdultMortality total number of outliers are 40.

c) For column NumInfantDeaths total number of outliers are 37.

d) For column Alcohol total number of outliers are 3.

e) For column PercentageExpenditure total number of outliers are 84.

f) For column HepB total number of outliers are 18.

g) For column Measles total number of outliers are 48.

h) For column BMI total number of outliers are 0.

i) For column Under5Deaths total number of outliers are 34.

j) For column Polio total number of outliers are 172.

k) For column Total.expenditure total number of outliers are 25.

l) For column Diphtheria total number of outliers are 170.

m) For column HIV total number of outliers are 69.

n) For column GDP total number of outliers are 69.

o) For column thinness1.19y total number of outliers are ‘r num_outliers.tthinness1.19y.

p) For column thinness5.9y total number of outliers are 57.

q) For column Schooling total number of outliers are 28.

1.3 / Data Preparation

Question 6 - Normalize all numeric columns using z-score standardization. Explain in your markdown what you are doing and why normalization is necessary.

To prepare data I will be developing a function to first obtain Z score of each value in every column of the dataset and then replace the outliers thus found with NA. Later I will impute the missing values with mean of each column.

This step is necessary because it will ensure that the data is not skewed and all the values are in same scale for seamless knn function.

```
total_missing <- sum(is.na(LifeExp.df.z.norm))
```

Question 7 - Add a new, derived feature to the dataframe called “disease” that is the sum of the columns “HepB”, “Measles”, “Polio”, and “Diphteria”.

```
# Adding Column named Disease
LifeExp.df.z.norm$Disease <- rowSums(LifeExp.df.z.norm[, c("HepB", "Measles", "Polio", "Diphtheria")], 
```

1.4 / Sampling Training and Validation Data

Question 8 - Data. Randomize (shuffle) the data and create a stratified sample where you randomly select 15% of each of the cases for each “status” column value to be part of the validation data set, so 15% of the “Developing” and 15% of the “Developed”. The remaining cases will form the training data set. Put the training and validation data into new dataframes.

The data was randomized and a stratified sample was developed, where 15% of each case for each “Status” type was randomly selected were included in the validation data set The remaining cases were used to form the training data set.

1.5 / Predictive Modeling

Apply the knn function from the class package with k=5 to predict the country status for the following new data point (you can impute the missing values/columns using median). Explain what you did (not not show the code), explain what algorithm you used to make the prediction and why this algorithm is useful, and then make a prediction. Make sure you standardize the new data values the same way as you standardized the training data or distance calculations will not be meaningful.

Life expectancy = 66.4 | Adult Mortality = 275 | infant deaths = 1 | Alcohol = 0.01 | percentage expenditure = 10| Hepatitis B = 40 |Measles = 400 | BMI = 17 | GDP = 620 | under-five deaths = 106 | Polio = 10 | Diphtheria = 66

To predict the outlook for the given new data points, I first developed a Data frame. Imputed the missing values using median. Standardized it with z score standardization method.

```r
# Load required libraries
library(class)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

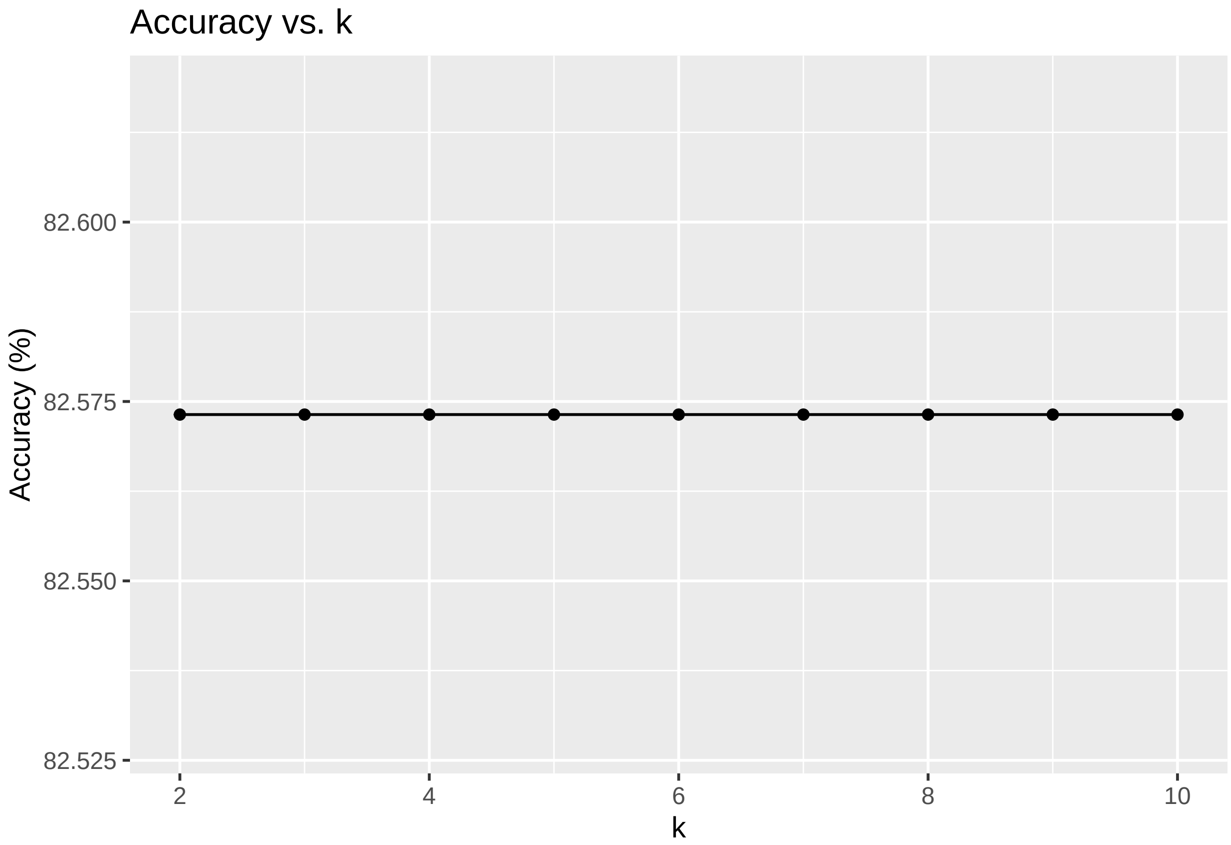
```r
# Load the dataset (replace 'data.csv' with your dataset's filename)
url <- "https://s3.us-east-2.amazonaws.com/artificium.us/datasets/LifeExpectancyData.csv"
data <- read.csv(url)

# Convert columns to appropriate types
numeric_columns <- c(
  "Year", "LifeExpectancy", "AdultMortality", "NumInfantDeaths",
  "Alcohol", "PercentageExpenditure", "HepB", "Measles", "BMI",
  "GDP", "Under5Deaths", "Polio", "Diphtheria"
)

data[numeric_columns] <- lapply(data[numeric_columns], as.numeric)

# Replace missing values with medians for numeric columns
data_filled <- data %>%
  mutate(across(all_of(numeric_columns), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))

# Standardize the data
mean_values <- colMeans(data_filled[, numeric_columns])
sd_values <- apply(data_filled[, numeric_columns], 2, sd)

data_standardized <- as.data.frame(scale(data_filled[, numeric_columns], center = mean_values, scale =

# Standardize the new data point
```

```r
new_data_point <- data.frame(
  Year = 2000,
  LifeExpectancy = 66.4,
  AdultMortality = 275,
  NumInfantDeaths = 1,
  Alcohol = 0.01,
  PercentageExpenditure = 10,
  HepB = 40,
  Measles = 400,
  BMI = 17,
  GDP = 620,
  Under5Deaths = 106,
  Polio = 10,
  Diphtheria = 66
)

new_data_standardized <- as.data.frame((as.matrix(new_data_point) - mean_values) / sd_values)

# Use k-NN algorithm to predict country status for the new data point
k <- 5
predicted_status <- knn(data_standardized, new_data_standardized, data_filled$Status, k)

# Display the predicted country status
cat("Predicted Country Status:", predicted_status, "\n")
```

```
## Predicted Country Status: 2
```

1.6 / Model Accuracy.

Question - 10 rom 2 to 10 versus accuracy (percentage of correct classifications). Display the chart and explain what you did and what it tells you. Through inspection of the plot, what value for k would you choose in your final model? Write all of this in markdown and do not echo the code for the chart. Again, make sure you standardize the new data values the same way as you standardized the training data or distance calculations will not be meaningful.



##2 / Predicting Shucked Weight of Abalones using Regression kNN (0 pts) Save the values of the "Shucked Weight" column in a separate vector called target_data and then also create a new dataset called train_data containing all the above training features (and, of course, not "Shucked Weight").

```
# Read the CSV file
url_Q2<- "https://s3.us-east-2.amazonaws.com/artificium.us/datasets/abalone.csv"

# obtain desired data
abalone.dframe<- read.csv(url_Q2, header = T, stringsAsFactors = F )
# Observe data
str(abalone.dframe)
```

```
## 'data.frame':    4178 obs. of  9 variables:
##  $ Length       : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
##  $ Diameter     : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
##  $ Height       : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
##  $ ShuckedWeight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
##  $ VisceraWeight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
##  $ ShellWeight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
##  $ WholeWeight  : num  0.514 0.226 0.677 0.516 0.205 ...
##  $ NumRings     : int  15 7 9 10 7 8 20 16 9 19 ...
##  $ Sex          : chr  "M" "M" "F" "F" ...
```

```
summary(abalone.dframe)
```

```
##      Length          Diameter          Height        ShuckedWeight
##  Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0010  
##  1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.1861  
##  Median :0.545   Median :0.4250   Median :0.1400   Median :0.3360  
##  Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.3595  
##  3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:0.5020  
##  Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :1.4880  
##  VisceraWeight     ShellWeight       WholeWeight        NumRings     
##  Min.   :0.0005   Min.   :0.0015   Min.   :0.0020   Min.   : 1.000  
##  1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.:0.4416   1st Qu.: 8.000  
##  Median :0.1710   Median :0.2340   Median :0.7997   Median : 9.000  
##  Mean   :0.1806   Mean   :0.2389   Mean   :0.8290   Mean   : 9.934  
##  3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:1.1538   3rd Qu.:11.000  
##  Max.   :0.7600   Max.   :1.0050   Max.   :2.8255   Max.   :29.000  
##      Sex           
##  Length:4178       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

```r
colSums(is.na(abalone.dframe))
```

```
##        Length      Diameter        Height ShuckedWeight VisceraWeight 
##             0             0             0             0             0 
##   ShellWeight   WholeWeight      NumRings           Sex 
##             0             0             0             0 
```

(0 pts) Save the values of the "Shucked Weight" column in a separate vector called target_data and then also create a new dataset called train_data containing all the above training features (and, of course, not "Shucked Weight").

```r
# Extract the 'Shucked Weight' column values into the 'target_data' vector
target_data_col <- abalone.dframe$ShuckedWeight
target_data <- data.frame(ShuckedWeight = target_data_col)
target_data <- as.numeric(target_data$ShuckedWeight)

# Create the 'train_data' dataset with all training features except 'ShuckedWeight'
train_data <- abalone.dframe[, -4]  # Exclude the 4th column (shuckedweight)
```

The abalone dataset is split into feature data frame and target data frame (ShuckedWieght).

#2.2 Encode all categorical columns using an encoding scheme of your choice but document (in markdown) why you chose it.

```r
# Convert values in Sex column to factors
sex <- as.factor(train_data$Sex)

# Obtain levels of factors present in this column
l <- length(levels(sex))
l
```

```
## [1] 3
```

With frequency encoding, the categorical variable can be converted to numeric variable and still be include in just one-column. Thus, it is a effective method for distance-based algorithms like kNN.

The "Sex" column is replaced with the frequencies of males 0.37, females 0.31, and infants 0.32 in the dataset.

##2.3 Normalize appropriate columns in train_data using min-max normalization.

#2.4, Build (write yourself) a function called knn.reg that implements a regression version of kNN that averages the value of the "Shucked Weight" of the k nearest neighbors using a weighted average where the weight is 3 for the closest neighbor, 2 for the second closest, and 1 for the remaining neighbors (recall that a weighted average requires that you divide the sum product of the weight and values by the sum of the weights).

Developed kNN function from scratch using Euclidean distance and given weights to obtain weighted average.

#5. Forecast the Shucked Weight of this new abalone using your regression kNN using k = 3: Sex: M | Length: 0.34 | Diameter: 0.491 | Height: 0.245 | Whole weight: 0.4853 | Viscera weight: 0.0887 | Shell weight: 0.19 | Rings: 10

The new data points given are normalized using min-max normalization method.

```
## Warning in nearest_shuckedweight * c(3, 2, rep(1, k - 3)): longer object length
## is not a multiple of shorter object length
```

#6.Calculate the Mean Squared Error (MSE) using a random sample of 15% of the data set as test data. The code needs to be carefully constructed for efficiency so that questions (5) and (6) have a reasonable chance of completing. Otherwise, it can take too long to run. Common problems are loops and the dynamic creation of data frames: pre-allocate memory and use which, vector calculations, and apply when possible, rather than loops. Use Sys.time to measure the run time of parts of code to determine bottlenecks.

```
## [1] 0.06510543
```

3 / Forecasting Future Sales Price

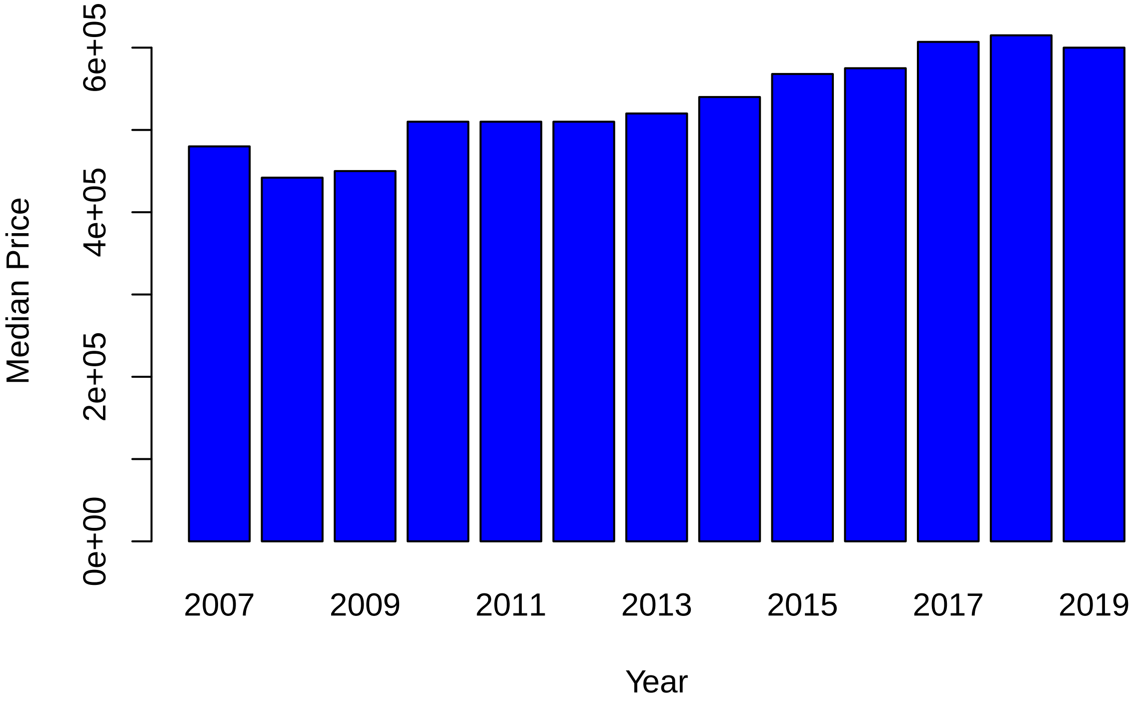
install.packages(tinytex) library(tinytex)

We obtained a data set with a total of r num_transactions sales transactions for the years r start_year to r end_year. The median sales price for the entire time frame was r format(round(median_price,0), big.mark = ",", scientific = FALSE) , while the 10% trimmed mean was r format(round(trimmed_mean_10,1), big.mark = ",", scientific = FALSE) (sd = r format(round(sd_price,0), big.mark = ",", scientific = FALSE))

Broken down by year, the following are the average sales prices per year:

```
## # A tibble: 13 x 3
##    year  trimmed_mean_10 median_saleprice
##    <chr>           <dbl>            <dbl>
##  1 2007          489189.           480000
##  2 2008          463226.           442000
##  3 2009          470796.           450000
##  4 2010          528650.           510000
##  5 2011          528367.           510000
##  6 2012          522673.           510000
##  7 2013          532317.           520000
##  8 2014          558258.           540000
##  9 2015          586370.           568000
## 10 2016          593631.           575000
## 11 2017          628977.          607000
## 12 2018          623654.          615000
## 13 2019          608294.          600000
```

As the graph below shows, the median sales price per year has been increasing. However, there has been a slight decrease in the last year.



Using a weighted moving average forecasting model that averages the prior 3 years (with weights of 4, 3, and 1), and a linear regression trend line model, we predict next year's average sales price to be around $r format(round(average_of_forecasts,0), big.mark = ",", scientific = FALSE).