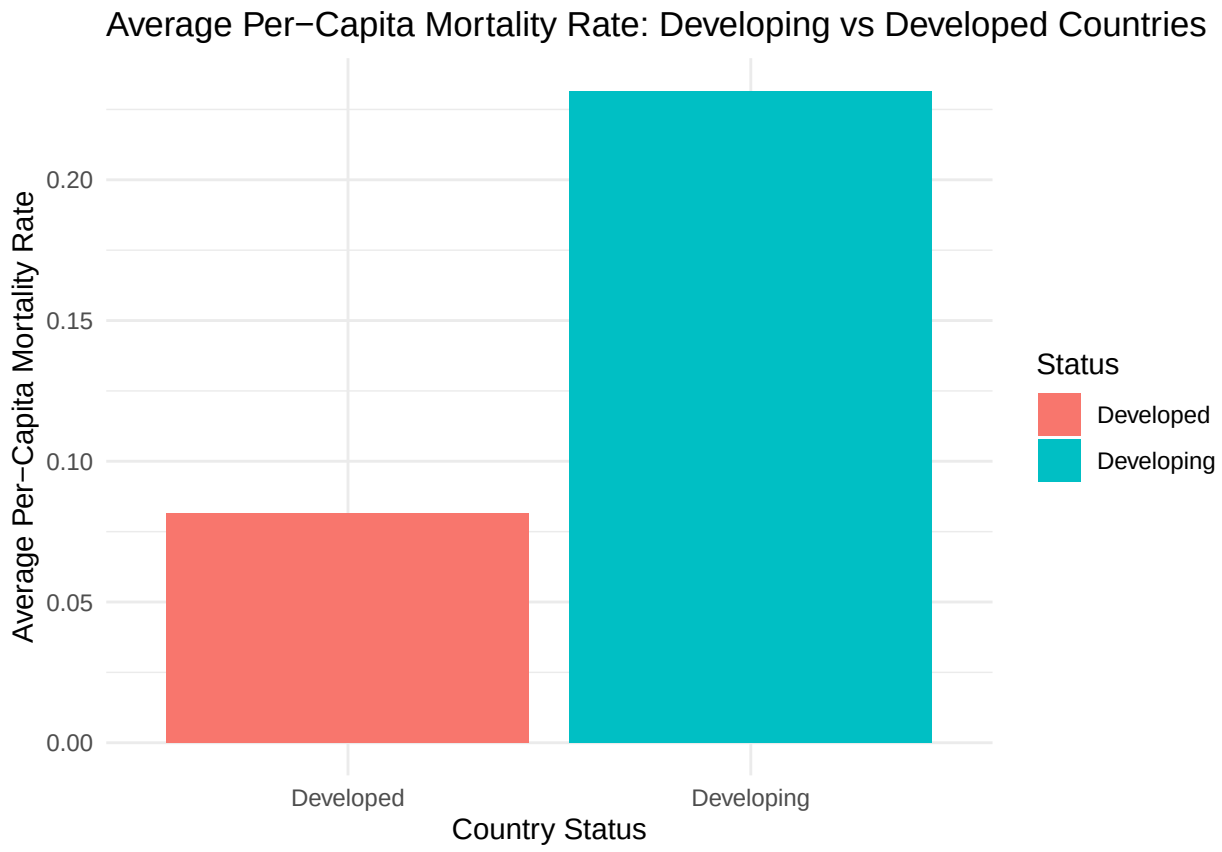
BUILD: Practicum 1

Dauby, Ray

09-30-24

1 / Predicting Life Expectancy

1.1 / Analysis of Per Capita Mortality



The per capita mortality rate of Developing countries is 2.85 times that of Developed countries. Individually, Developed countries have a per-capita mortality rate of 0.08138, while Developing countries have a per-capita mortality rate of 0.23163.

Statistical Significance of Mortality

- The difference in mortality between the two types of countries is statistically significant, because the p-value extracted from the significance test is $2.5136813 \times 10^{-88}$. The Shapiro-Wilk test was used to test the normality of the data (Using cutoffs of 0.9 for W and 0.05 for p), which was shown to be not

normally distributed. Because the data was not normally distributed, the Wilcoxon Rank-Sum Test was used to assess the significance of the difference in mortality rate per capita between the groups of Developed and Developing countries.

Normality of Life Expectancy Data

- The Shapiro-Wilk test yielded a p-value of $7.3604623 \times 10^{-29}$, which is lower than the conventional significance level of 0.05. Additionally, the W-statistic for the distribution of life expectancy is greater than 0.9, indicating a normal distribution. These results indicate that we should accept the null hypothesis, which states that the data is normally distributed.

1.2 / Identification of Outliers

Column	Number_of_Outliers	PercentOutliers
row	172	NA
Life.expectancy	5	0.1708
Adult.Mortality	52	1.776
infant.deaths	50	1.702
Alcohol	5	0.1822
percentage.expenditure	97	3.302
Hepatitis.B	159	6.667
Measles	51	1.736
under.five.deaths	36	1.225
Polio	172	5.892
Total.expenditure	28	1.032
Diphtheria	170	5.824
HIV.AIDS	73	2.485
GDP	88	3.534
Population	18	0.7874
thinness..1.19.years	66	2.273
thinness.5.9.years	71	2.445
Income.composition.of.resources	130	4.691
Schooling	28	1.009

The table above contains the numbers of outliers present in each numeric column of the dataset. Outliers are defined as more than 2.8 standard deviations from the mean, and were identified using Z-scoring. Outliers can be handled in several ways, notably we can A) trim the dataset to remove outliers, B) scale the data to better fit our model using a log or square root transformation, or C) replace outliers with imputed values such as the mean or median.

The maximum life expectancy was 89, and the minumum life expectancy was 36.3, with a standard deviation of 9.5 and a median of 72.1. Calculating a trimmed mean is not a good option with the life expectancy data, because based on the outlier counts, there are not very many and their effect on the central tendencies of the data is relevant and should not be removed. In order to calculate the percent I would trim, I would identify what percent of the data is made up of outliers, and trim about that much off either tail. In practice, the life expectancy column has very few outliers, and given the nature of this data, I do not think they should be removed.

1.3 / Data Preparation

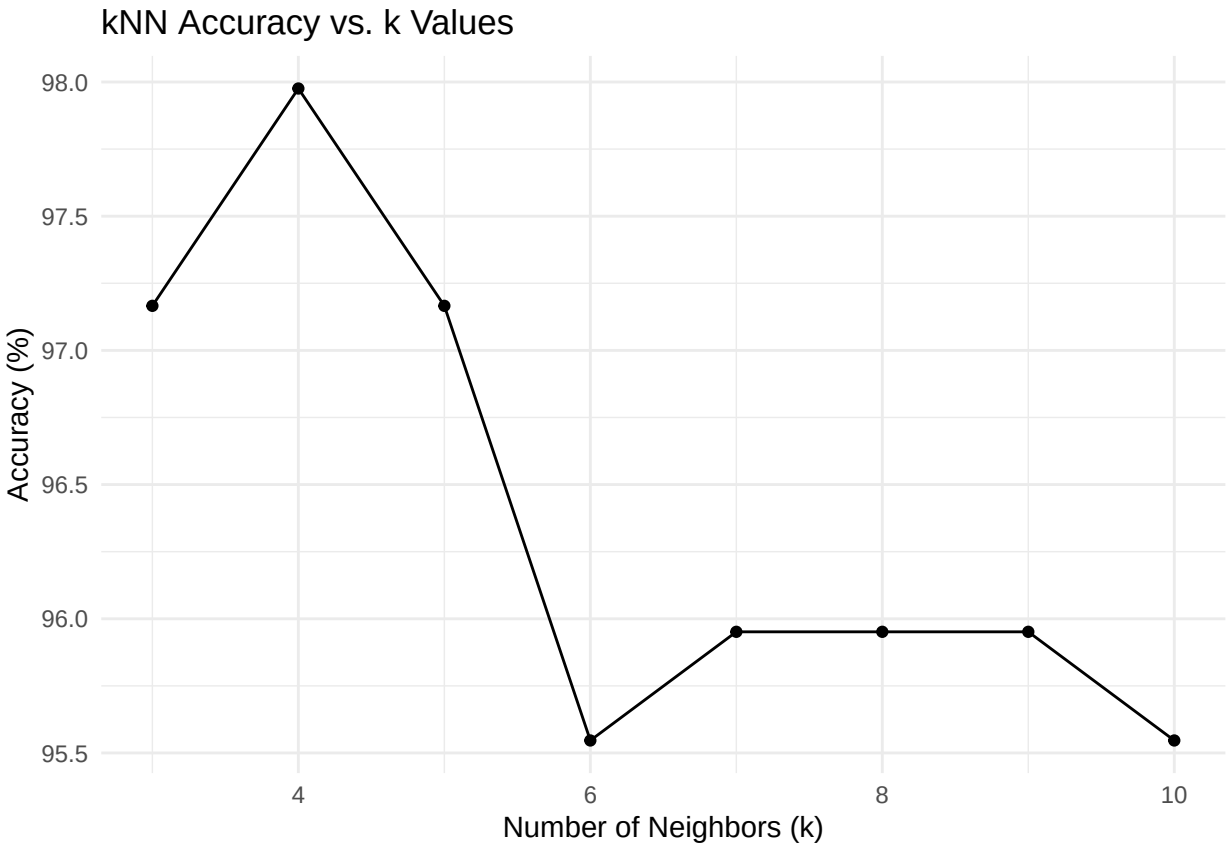
Normalizing the data using Z-scoring ensures that the data for each feature is on a similar scale, which is important when dealing with algorithms that are sensitive to scale, such as distance-based algorithms (k-nearest neighbors).

1.4 / Sampling Training and Validation Data

1.5 / Predictive Modeling

The kNN function from the class package with k = 6 was applied to predict the country status for the data point above. Missing values were imputed using the median of the original dataset. The new data point was standardized using z-scoring, as were the training and test sets. The model predicts that the new data point is a Developing country. The kNN model classifies new data points based on the majority class of their k nearest neighbors in the feature space, which makes it useful for classification tasks such as this one.

1.6 / Model Accuracy



The plot above shows the percent accuracy of the trained model on the testing data using values of k from 3 - 10. I used the kNN model that I wrote to test the ccuracy of the model and identify the best value of k to use with my data. Through inspection of the plot, I would choose k = 4 for my final model, as it reports the highest accuracy.

2 / Predicting Shucked Weight of Abalones using Regression kNN

2.1 / Creating Training and Target Datasets

The Training Set (Example Rows):

```
##   Length Diameter Height VisceraWeight ShellWeight WholeWeight NumRings Sex
## 1  0.455    0.365  0.095        0.1010       0.150      0.5140       15   M
## 2  0.350    0.265  0.090        0.0485       0.070      0.2255        7   M
## 3  0.530    0.420  0.135        0.1415       0.210      0.6770        9   F
## 4  0.440    0.365  0.125        0.1140       0.155      0.5160       10   F
## 5  0.330    0.255  0.080        0.0395       0.055      0.2050        7   I
```

2.2 / Encoding Categorical Variables

I chose to encode the categorical column "Sex" using the one-hot scheme, which creates an explicit column of binary data for each category (M/F/I) that is easily accessed by the model. I am using min-max normalization for the numeric data, so the binary encoded data will be on the same scale as the rest of the dataset.

2.3 / Min-Max Normalization

2.4 / Writing kNN Function

2.5 / Forecasting Shucked Weight

The predicted weight for the provided data point is 0.221.

2.6 / Calculating Mean Squared Error

```
## Time taken for sampling:  0.002228022
```

```
## Time taken for predictions:  3.09094
```

```
## Mean Squared Error (MSE): 0.002038888
```

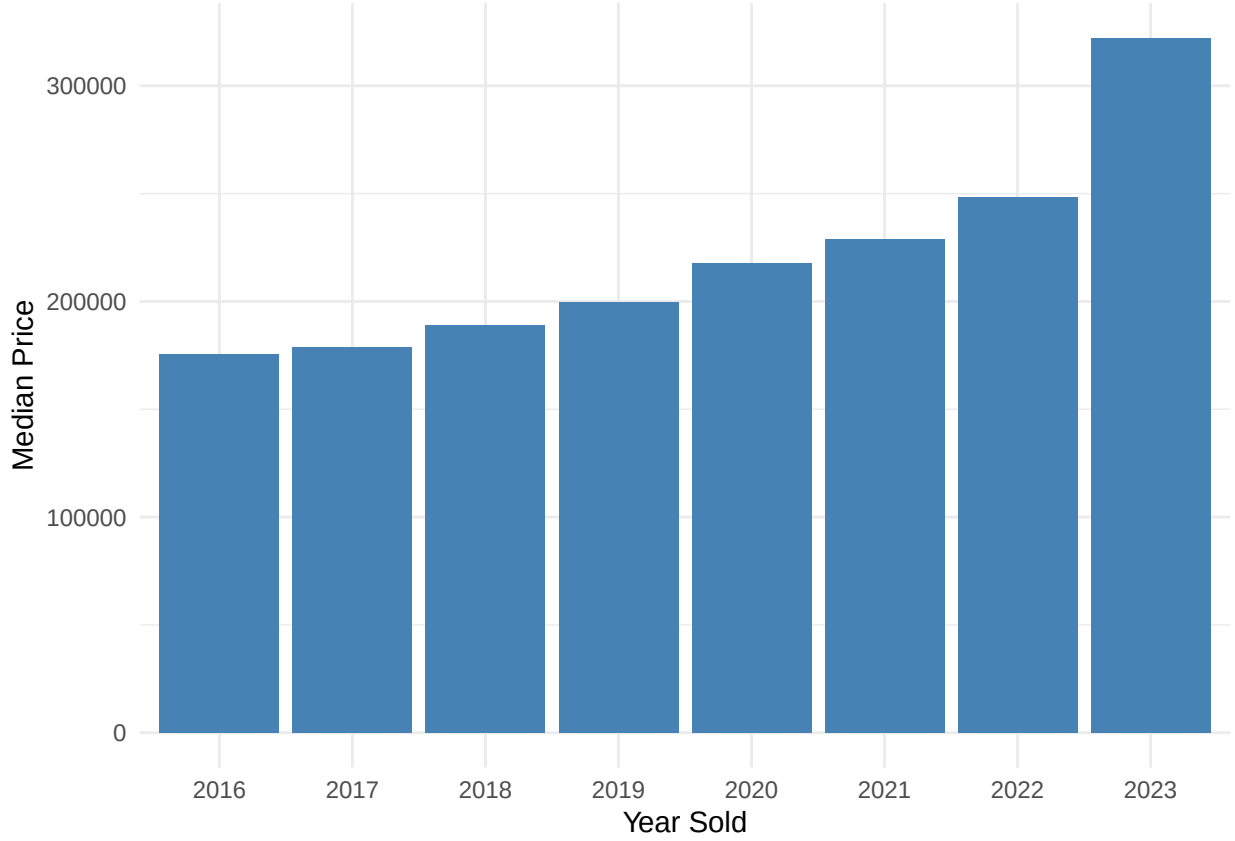
3 / Forecasting Future Sales Price

We obtained a data set with a total of 100 sales transactions for the years from 2016 to 2023. The median sales price for the entire time frame was $207000, while the 10% trimmed mean was $212869 (sd = $72614). Broken down by year, we have the following number of sales, plus the 10% trimmed mean and median sales prices per year:

YearSold	TotalHomesSold	TrimmedMeanPrice	MedianPrice
2016	6	169170.0	175500
2017	18	189348.8	178965
2018	20	180810.0	188910
2019	8	201037.5	199800
2020	9	220500.0	217800

YearSold	TotalHomesSold	TrimmedMeanPrice	MedianPrice
2021	12	230562.0	229050
2022	13	252327.3	248400
2023	13	310909.1	322200

As the graph below shows, the median sales price per year has been increasing.



Using both a weighted moving average forecasting model that averages the prior 3 years (with weights of 0.7, 0.2, and 0.1) and a linear regression trend line model, we predict next year's average sales price to be around $299615 (average of the two forecasts). The average home price of homes with both pools and air conditioning changed from $218700 in 2017 (earliest year for which data is available) to $624600 in 2023 (most recent year's sales data).