Tiance Tan
Jacek Plonowski

Homework 1

Cereal Data Analysis and Storyboard:

1. List five analytics queries or questions that you would have about this dataset in your exploratory process.

1) Do any of the data indicate which cereals will be the healthiest option to provide for students?

2) Do any of the data indicate who are their best customers (age, weight difference)?

3) Do any of the data help forecast which cereal is more likely to be better received by the students of the school?

4) Do any of the data reveal which manufacturer tends to deliver a better product, and thus is better to do business with?

5) Does the rating correlate with any other columns such as calories, protein, etc?

2. Describe 5 visualizations that can help you explore this data. (Example: line graph between variable x and y).

1) Stacked/unstacked bar chart of amount of protein/fat/sugar/sodium. I.e. nutritional data people use to judge food.

2) Scatter plot of the ‘rating’ and ‘shelf’ values for each cereal, analyze potential correlation between the location of the item in the store and its rating.

3) Box/Whisker plot to show distributions of ‘rating’ , subsetting the dataset for each manufacturer. E.g. compare ratings for Post cereals to Quaker Oats cereals, which one has generally better rated cereal, etc.

4) Bar chart for each manufacturer and its respective mean product rating

5) Pie chart for each manufacturer. Help us to find which manufacturer has the largest proportion of the cereal market.

3. What other data would be helpful to prepare your presentation? What do you think is your boss's goal?

Another useful piece of information would be some sort of data related to the pricing/costs associated with each manufacturer/product. We are assuming that besides the healthiness and rating of the product, there are other factors that influence which option will be chosen, such as cost/availability/potential discounts, etc. Meanwhile, it would be great to have the data related to the supply level for each product. Since the school may require lots of cereal everyday, we have to ensure that the manufacturer can produce a large amount of cereal.

We think our boss's goal is to find the cereal that provides the best overall nutrition also with a good rating.

4. Assume that after a long exploratory data analysis you reach a recommendation for your boss (make assumptions as needed). Provide a one-sentence "big idea."

Big Idea:

We recommend Nabisco to be the manufacturer responsible for supplying breakfast cereals for our school, specifically the "Shredded Wheat 'n' Bran" cereal, because of its positive reception among customers, good balance of nutrition content, and origin from a top performing manufacturer in the industry, providing the school with a healthy, popular, and reputable breakfast choice.

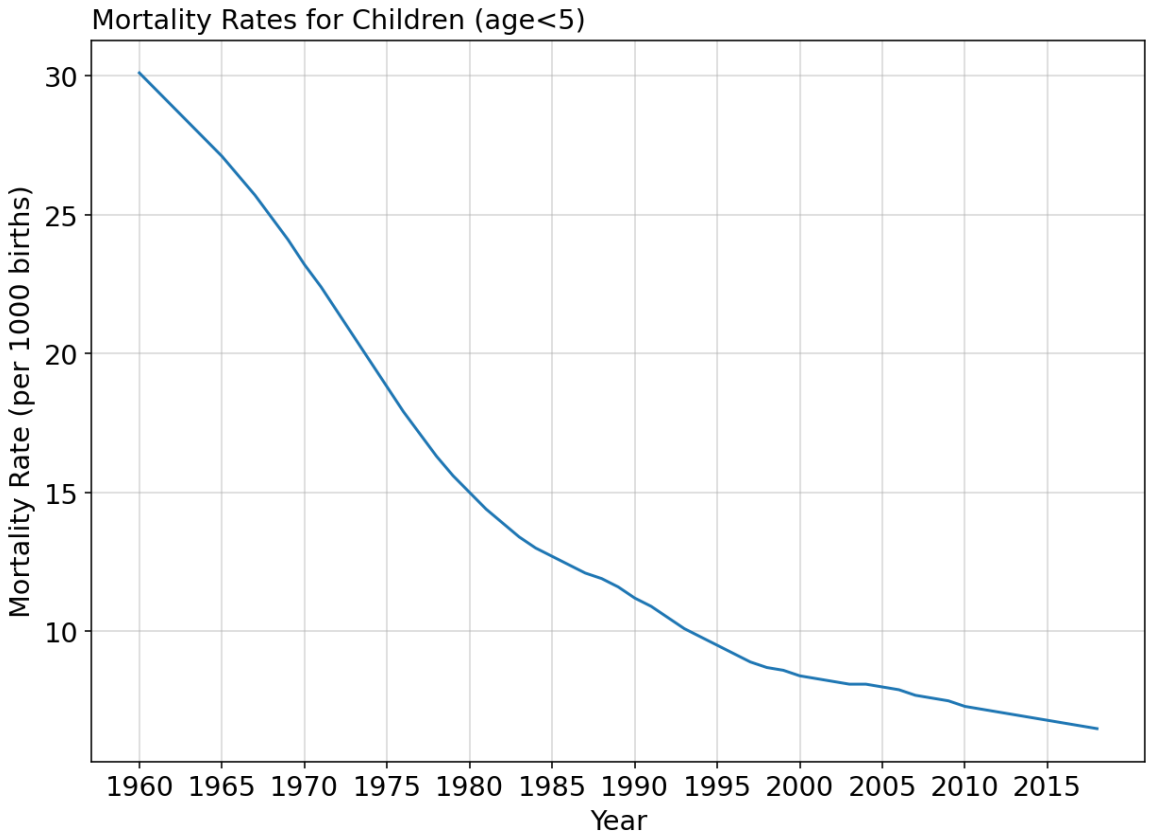
5. Create a storyboard for your presentation.

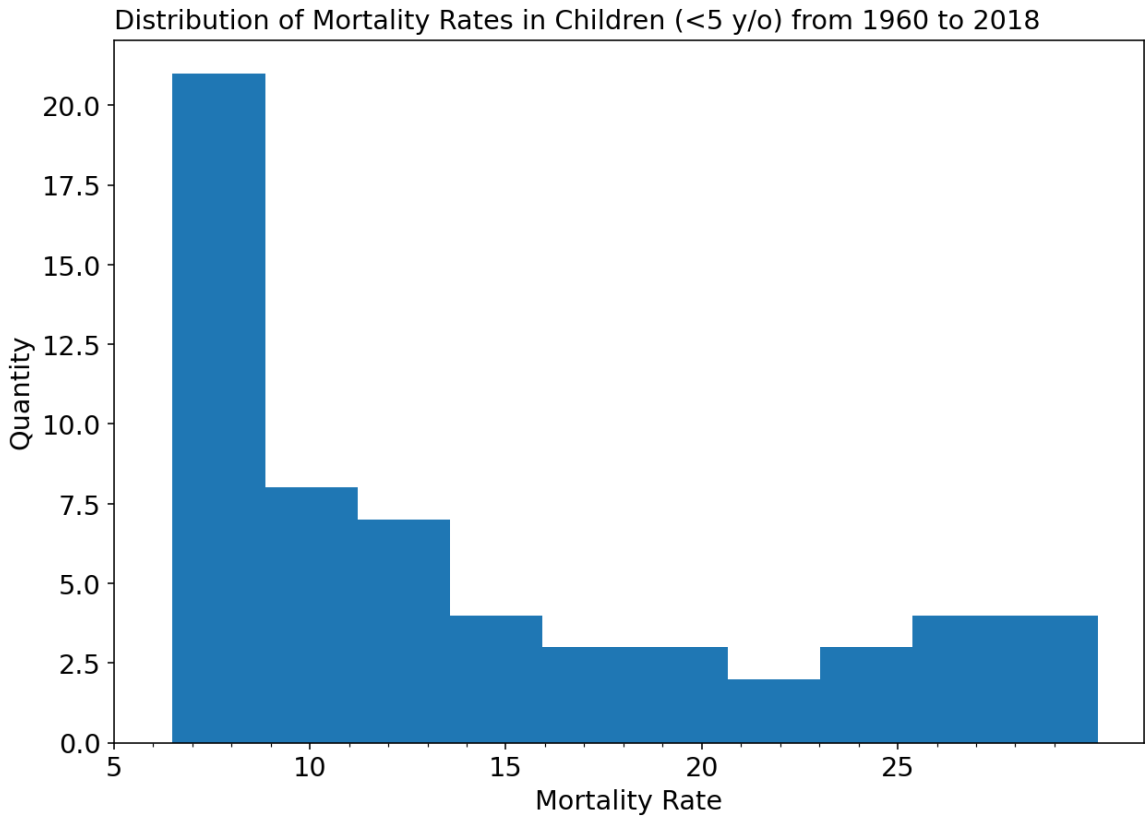
Introduction & The goal we are trying to address	Introduce 5 plots to help make a decision on cereal	1. Stacked/unstacked bar chart of nutrition facts. Help us determine the nutrition facts people are always looking for.
2. Scatter plot of the 'rating' and 'shelf' values for each cereal. Help us find people's preference on cereal brands and categories.	3. Box/Whisker plot to show distributions of 'rating'. Help us determine cereals with high ratings.	4. Bar chart for each manufacturer and its respective mean product rating. Help us choose manufacturers with great ratings.
5. Pie chart for each manufacturer. Choose large manufactures to ensure they can provide enough supply for school	Summary and Recommendation	

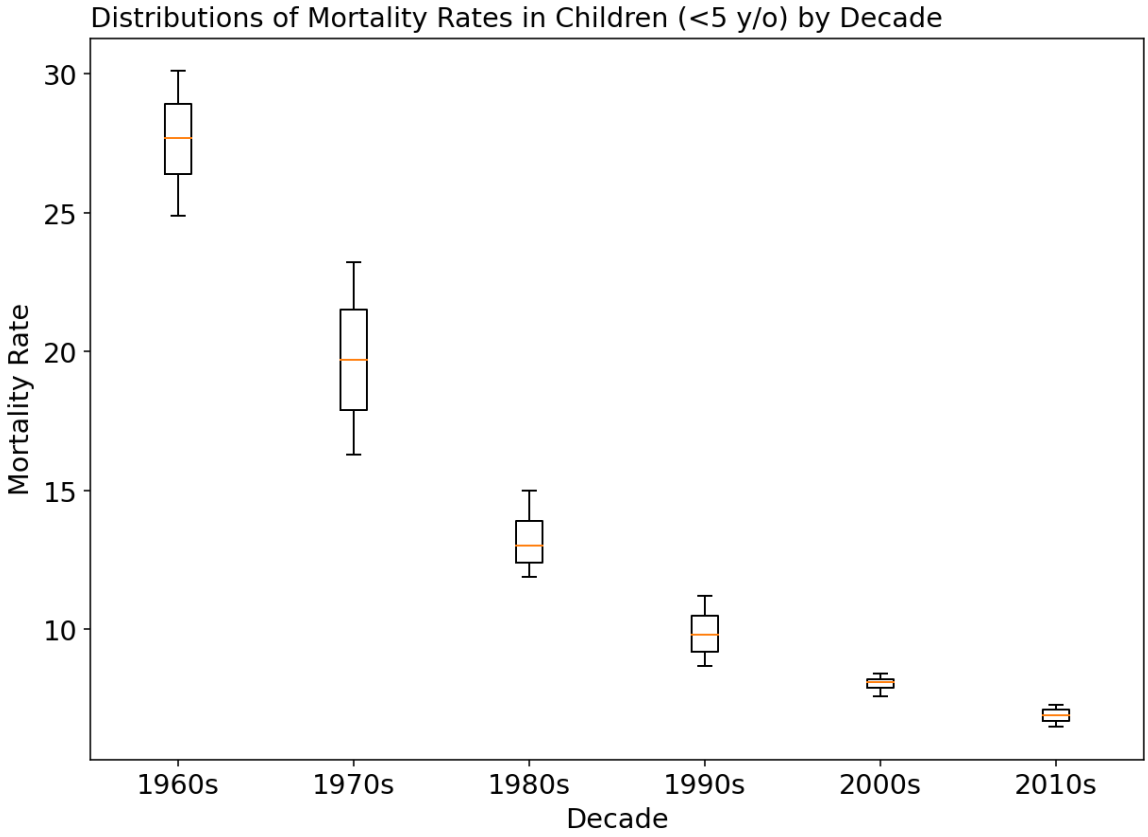
Practice Basic Matplotlib:

Practice using matplotlib to plot the mortality rate over time of children under 5 (per 1000 live births). Submit 3 visualizations for this data. Which one do you think works the best?

The plot that is the most effective in conveying the most important information in the dataset is probably the first plot, a line graph plotting the mortality rate values for each year from 1960 to 2018. The second visualization is a non-conventional look at the data, as it would help someone who was interested in what rates of mortality were most common during the years from 1960-2018. The third visualization is about as effective as the first plot, however, it has a more specialized use-case. It would aid someone who was interested in looking at the data decade by decade and drawing conclusions from the distribution of mortality rates within each decade.




Distribution of Mortality Rates in Children (<5 y/o) from 1960 to 2018
Quantity
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
5
10
15
20
25
Mortality Rate


Distributions of Mortality Rates in Children (<5 y/o) by Decade
Mortality Rate
10
15
20
25
30
1960s
1970s
1980s
1990s
2000s
2010s
Decade

Practice Slopegraph with Matplotlib: