Final Project Written Report

Introduction

The primary dataset belongs to the Centers for Disease Control and Prevention (CDC) and contains data regarding 5 common foodborne and waterborne diseases found across the USA. The 5 diseases of interest are vibrio, shiga toxin-producing escherichia coli (STEC), shigella, salmonella, and campylobacter. It is known that for the common flu, wintertime is typically the highest rate of transmission, and the highest number of cases are during those months. Here, the primary question will be to see if these foodborne and waterborne diseases follow a similar trend if they are more prevalent when the temperatures are low or if they are more prevalent when the temperature is higher and will be measured by the number of isolates of the diseases. Isolates are the number of cases for that particular disease. Using this information, we will be able to ascertain whether the food and water borne illnesses are more prevalent during the colder or warmer times of the year. Various cities across the United States will be utilized in this study to see the potential effects of geographical areas in foodborne disease cases. In addition, for exploratory purposes, the summary statistics of the 5 pathogens will be looked at as well to see if there is a particular pathogen that causes more cases than others and how spread out all 5 of them are, in terms of isolates.

Methods

In the CDC dataset, there are 10 different cities where data was collected across various regions in the mainland United States, ranging from the east coast to the west coast and northern cities to southern cities. The importance of having a wide range of cities in this study is to eliminate potential confounders that can arise from different geographical areas due to population size, average temperature, cultural, and others. For example, foodborne illnesses are more commonly found in seafood so one may say that a city like Boston, where people typically eat more seafood than in a place such as Kansas City. The cities were all accumulated into the average and median temperature and isolate values due to this reasoning.

The monthly temperature values were gathered from The National Oceanic and Atmospheric Administration (NOAA) and produced onto an excel sheet which was later merged into the CDC dataset based on the city (HHS_Region) and month. In the original dataset, the temperature values were collected very regularly, almost daily, for every city. Due to a very negligible to no deviation in temperatures and unfinished 2023 data, since the year is not yet over, the average temperatures were gathered for the month, regardless of the year being 2022 or 2023. Negligible in this scenario means that there were very little differences and essentially the same temperatures by day in 2022 and 2023. For example, January 22nd, 2022, in a city could be 33.1 degrees while in 2023 on the same date, it would be 33.3 degrees. This was a common feature across various randomly checked dates in numerous cities, therefore, due to the extremely minor deviation in temperature in 2022 and 2023, the monthly averages were conducted based off of the available data in those 2 years, without specifying the year. The seasonal temperature for each city was calculated based off the 4 main seasons with the months being December, January, and February for Winter, March, April, and May for Spring, June, July, and August for Summer, and September, October, and November for the Fall season. This dataset was used to analyze both the overall data by season and individual city by season.

The data was checked to make sure there were no missing values found in any of the variable categories. The dimensions, footers, and headers were all verified using various data cleaning and checking functions. The variable types made sense and were not changed and were assessed with the str, summary, and table functions. All the values in temperature, isolates, and months/quarters were plausible and made sense. The variable "Past two years average" for isolates were not utilized in this assignment as the values were not making sense because if you calculated for the averages for the same month, pathogen, and city, it would not equate to the correct calculated values from 2022-2023 given in the dataset. It appears it might have been calculated for 2021-2022 perhaps, but due to the lack of confirmation, it was not utilized in this study.

Results

To explore the relationships between temperatures & months and isolates & months, the summary statistics were calculated for those two. In the monthly temperatures table, the mean temperature values were used instead of median due to there being a normal distribution of the values. In the monthly isolates table, the median was preferred over the mean since the distribution was very skewed, as a result of some very high values resulting in an extremely high standard deviations in most months. Along with the summary statistic tables provided, histograms were created for both of these measures to ensure whether or not the data was normally distributed or not.

The results for the monthly temperature obviously showed that the highest average temperatures were in the summertime, with June, July, and August being the 3 highest temperature months and the 3 lowest temperature months being December, January, and February. Interestingly enough, the results for the number of pathogen isolates per month in terms of highest and lowest, were the exact same as the average temperature months. This would mean that the majority of pathogens spread during the summertime and are more stagnant towards the wintertime and cooler months of the year.

The seasonal results for each city showed a similar trend when starting from spring and ending with the winter season as it starts off with a gradual rise towards summer and once summer is over, the temperatures drop drastically until winter. Although, the temperature ranges vary per city, the overall trend maintains the same. When the seasonal data was done as an accumulation of all the cities represented per season per graph, it was relatively similar barring some cities in the Winter season, cities like San Francisco, Atlanta, and Dallas stood out compared to others with much higher temperatures on average. To answer the side question that was mentioned in the introduction, the number of isolates by pathogen were assessed by running a summary statistic on them. It showed that salmonella was, by far, the most common pathogen that was spread in this dataset and vibrio being the least common one, in terms of the average amount of pathogens per month across the various cities.

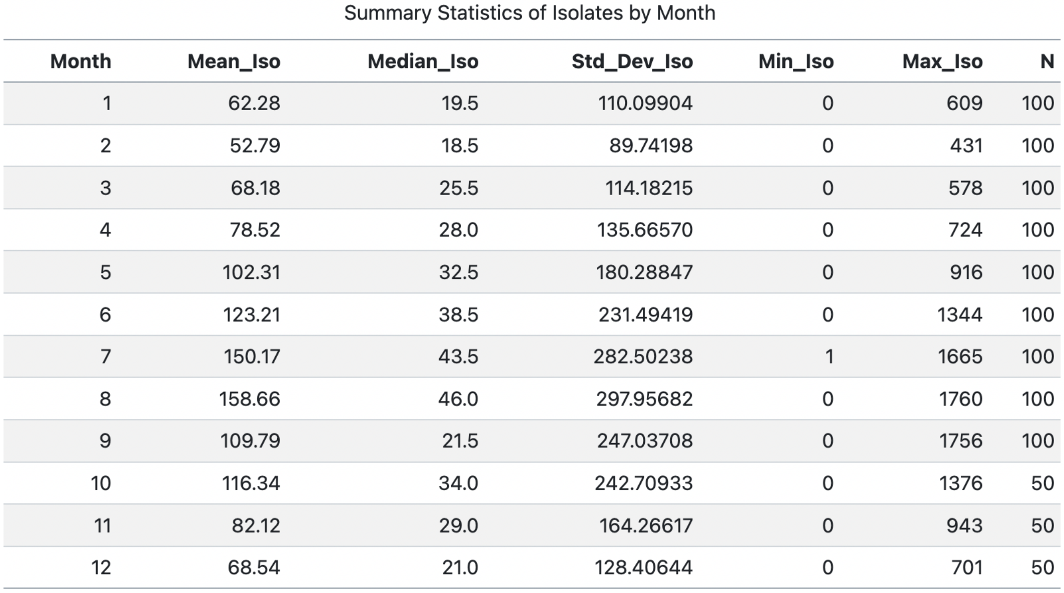
Bar graphs were the best form of visual representation for all 3 of these tables that were generated. It was interesting to see the similarity of the two graphs, depicting the isolate values and temperature values, both having a similar trend with the middle months having the highest temperatures and the early and late months having the lowest temperatures. Line graphs were utilized for the individual city reports for the 4 seasons to view the overall trend across the year.

Conclusion and Summary where you describe your findings.

In conclusion, unlike the flu, where the highest cases appear in the wintertime (lowest temperature months), the 5 common food and water borne illnesses that were utilized in this study show that the highest number of cases are in the summertime (highest temperature months). This would suggest that those pathogens spread, grow, and survive better in warmer temperatures than colder months across all the cities, on average. Salmonella was by far the most common type of pathogen in this dataset.

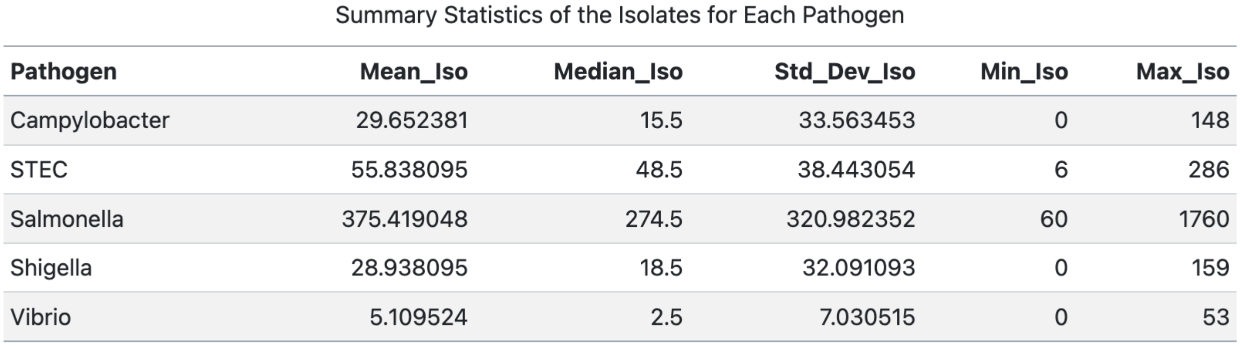
Final Project Written Report

Tables



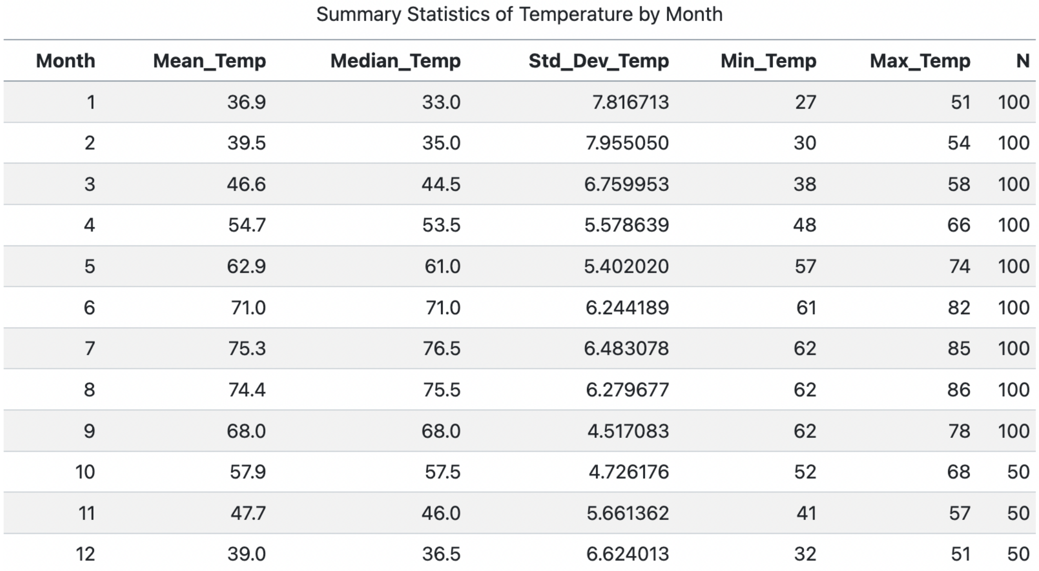
Summary Statistics of Isolates by Month

Month	Mean_Iso	Median_Iso	Std_Dev_Iso	Min_Iso	Max_Iso	N
1	62.28	19.5	110.09904	0	609	100
2	52.79	18.5	89.74198	0	431	100
3	68.18	25.5	114.18215	0	578	100
4	78.52	28.0	135.66570	0	724	100
5	102.31	32.5	180.28847	0	916	100
6	123.21	38.5	231.49419	0	1344	100
7	150.17	43.5	282.50238	1	1665	100
8	158.66	46.0	297.95682	0	1760	100
9	109.79	21.5	247.03708	0	1756	100
10	116.34	34.0	242.70933	0	1376	50
11	82.12	29.0	164.26617	0	943	50
12	68.54	21.0	128.40644	0	701	50



Summary Statistics of the Isolates for Each Pathogen

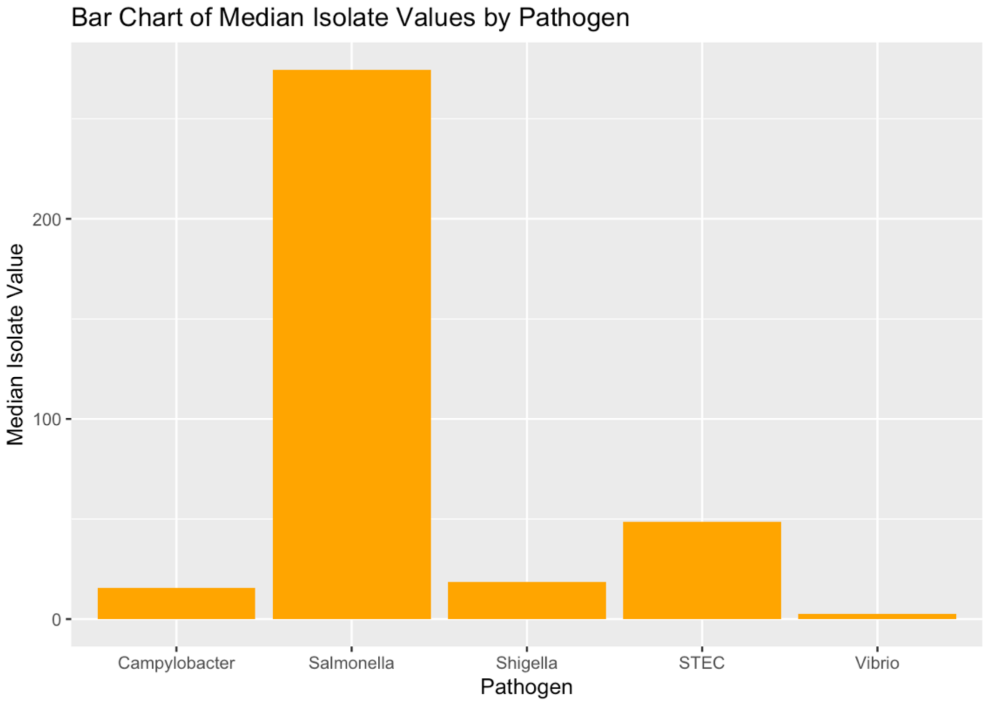
Pathogen	Mean_Iso	Median_Iso	Std_Dev_Iso	Min_Iso	Max_Iso
Campylobacter	29.652381	15.5	33.563453	0	148
STEC	55.838095	48.5	38.443054	6	286
Salmonella	375.419048	274.5	320.982352	60	1760
Shigella	28.938095	18.5	32.091093	0	159
Vibrio	5.109524	2.5	7.030515	0	53

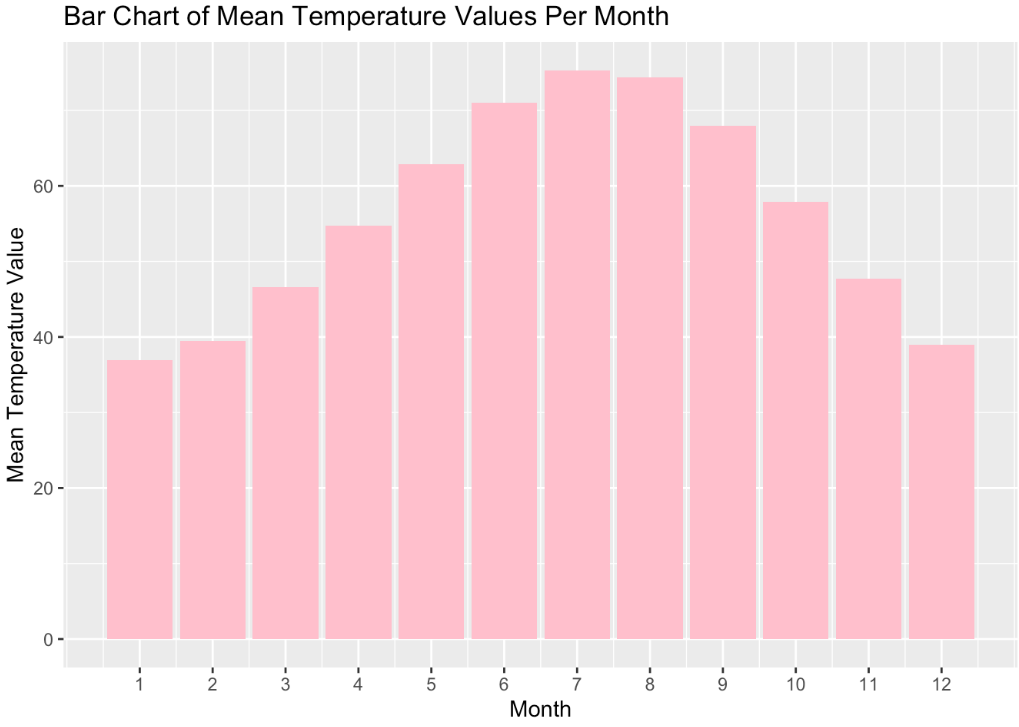


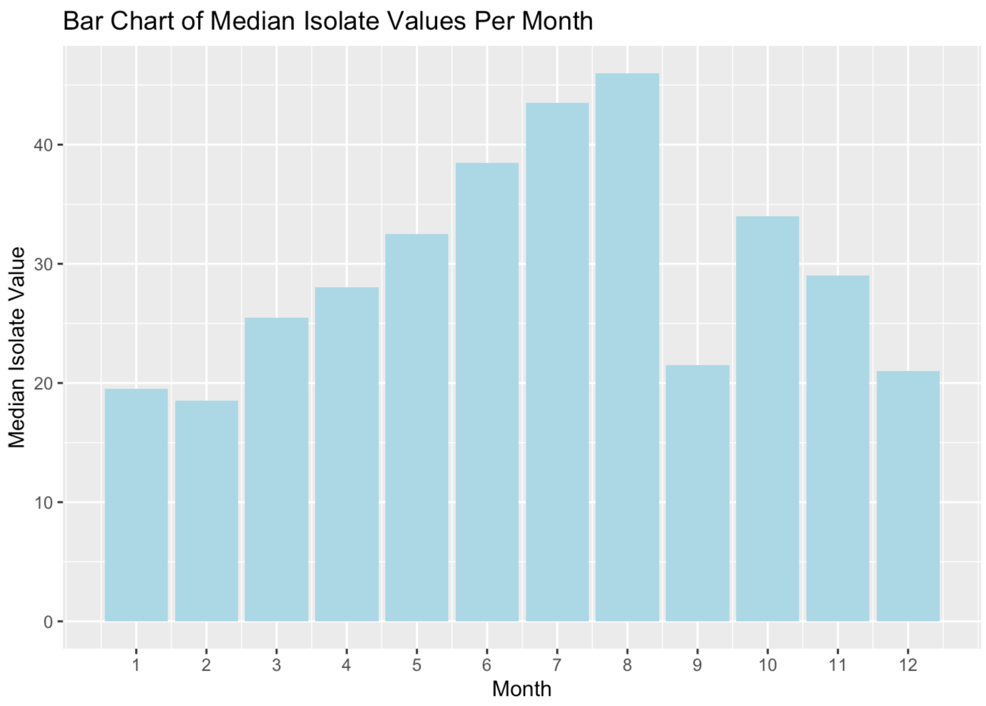
Summary Statistics of Temperature by Month

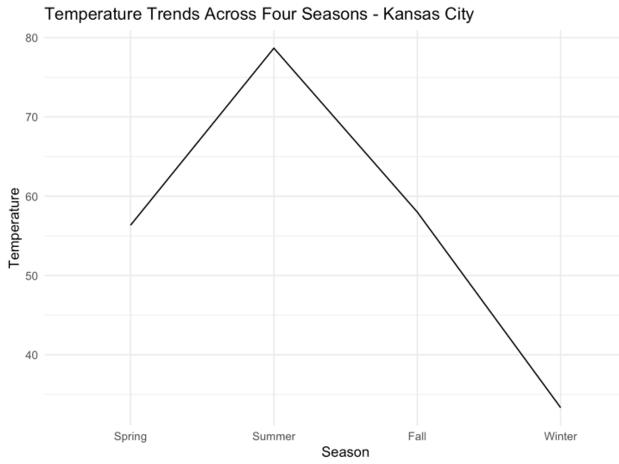
Month	Mean_Temp	Median_Temp	Std_Dev_Temp	Min_Temp	Max_Temp	N
1	36.9	33.0	7.816713	27	51	100
2	39.5	35.0	7.955050	30	54	100
3	46.6	44.5	6.759953	38	58	100
4	54.7	53.5	5.578639	48	66	100
5	62.9	61.0	5.402020	57	74	100
6	71.0	71.0	6.244189	61	82	100
7	75.3	76.5	6.483078	62	85	100
8	74.4	75.5	6.279677	62	86	100
9	68.0	68.0	4.517083	62	78	100
10	57.9	57.5	4.726176	52	68	50
11	47.7	46.0	5.661362	41	57	50
12	39.0	36.5	6.624013	32	51	50

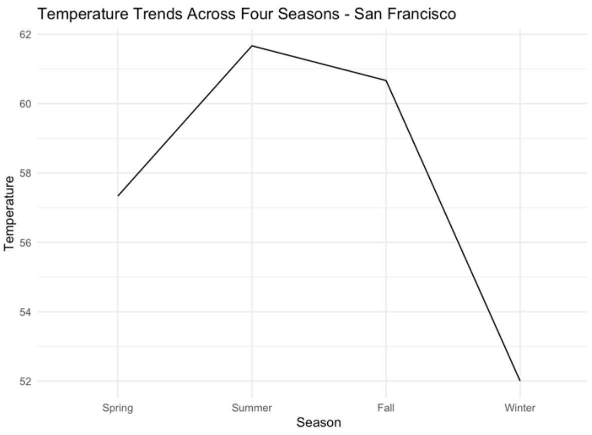
Figures

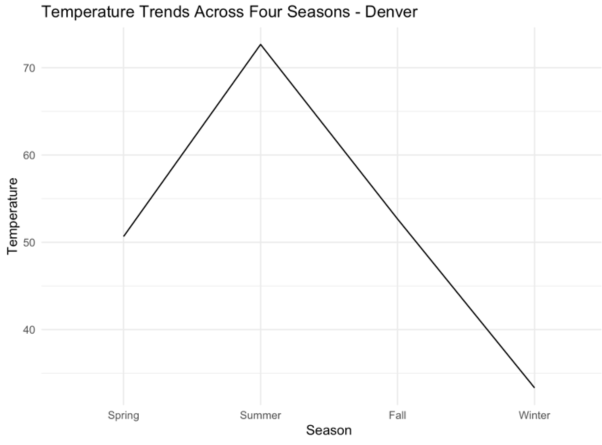

Bar Chart of Median Isolate Values by Pathogen
Median Isolate Value
200
100
0
Campylobacter
Salmonella
Shigella
STEC
Vibrio
Pathogen

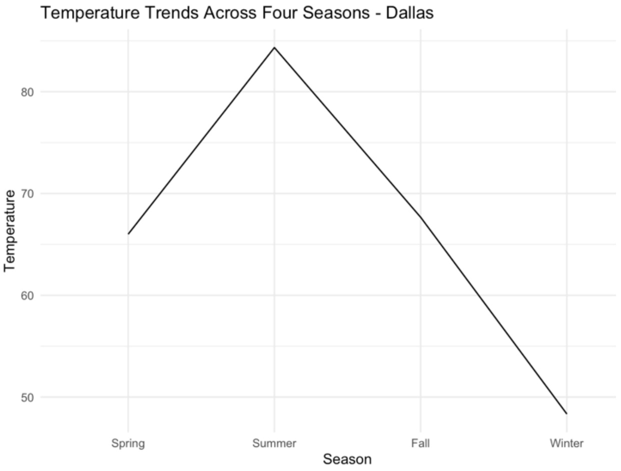

Bar Chart of Mean Temperature Values Per Month
Mean Temperature Value
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

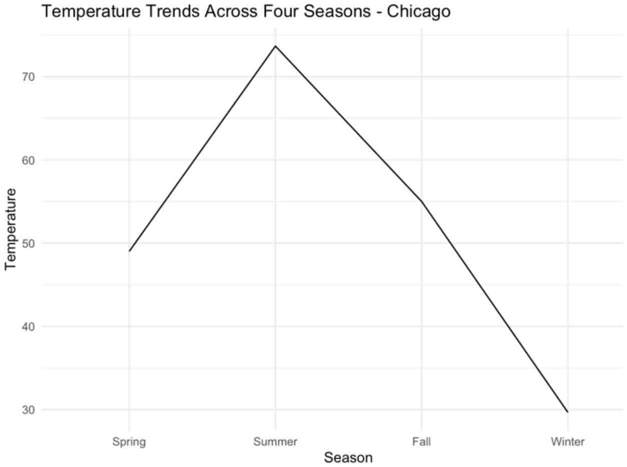

Bar Chart of Median Isolate Values Per Month
Median Isolate Value
0
10
20
30
40
Month
1
2
3
4
5
6
7
8
9
10
11
12

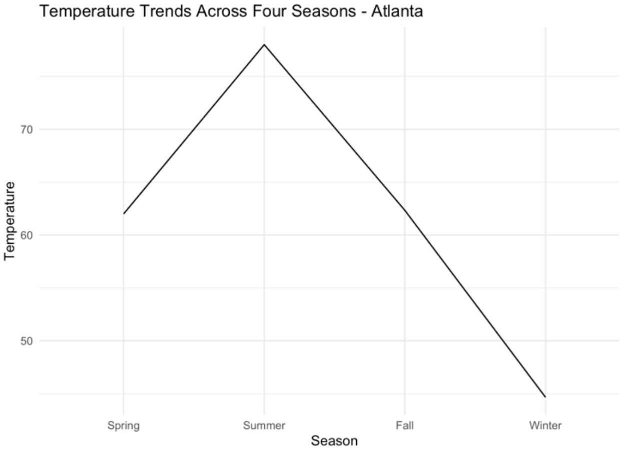

Temperature Trends Across Four Seasons - Kansas City
Temperature
40
50
60
70
80
Season
Spring
Summer
Fall
Winter

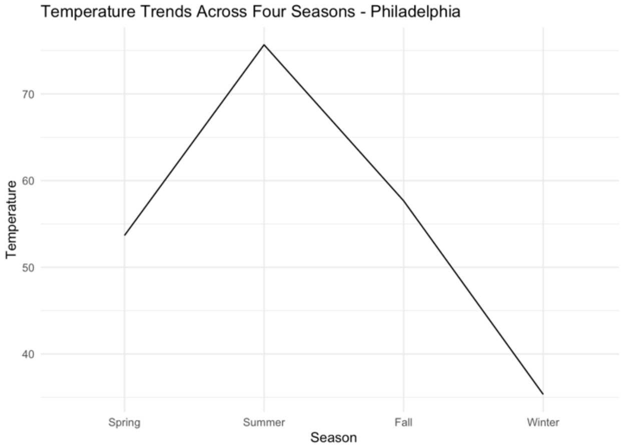

Temperature Trends Across Four Seasons - San Francisco
Temperature
52
54
56
58
60
62
Season
Spring
Summer
Fall
Winter

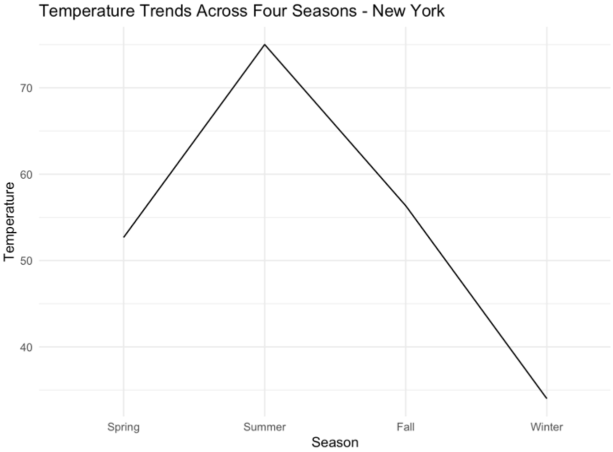

Temperature Trends Across Four Seasons - Denver
Temperature
40
50
60
70
Season
Spring
Summer
Fall
Winter

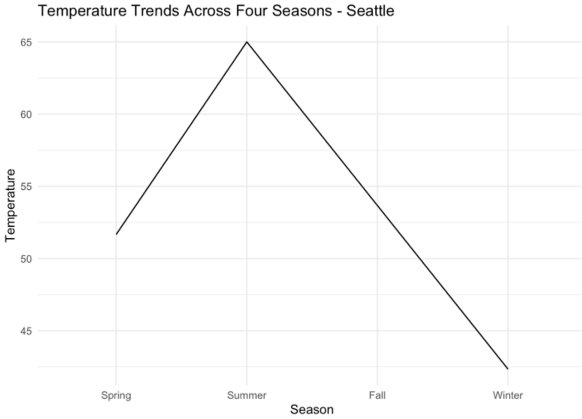

Temperature Trends Across Four Seasons - Dallas
Temperature
50
60
70
80
Season
Spring
Summer
Fall
Winter

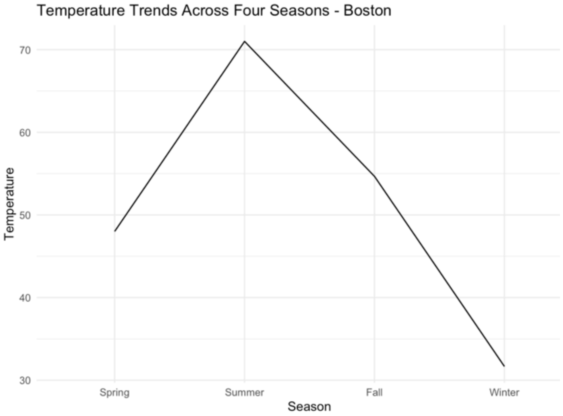

Temperature Trends Across Four Seasons - Chicago
Temperature
70
60
50
40
30
Spring
Summer
Fall
Winter
Season

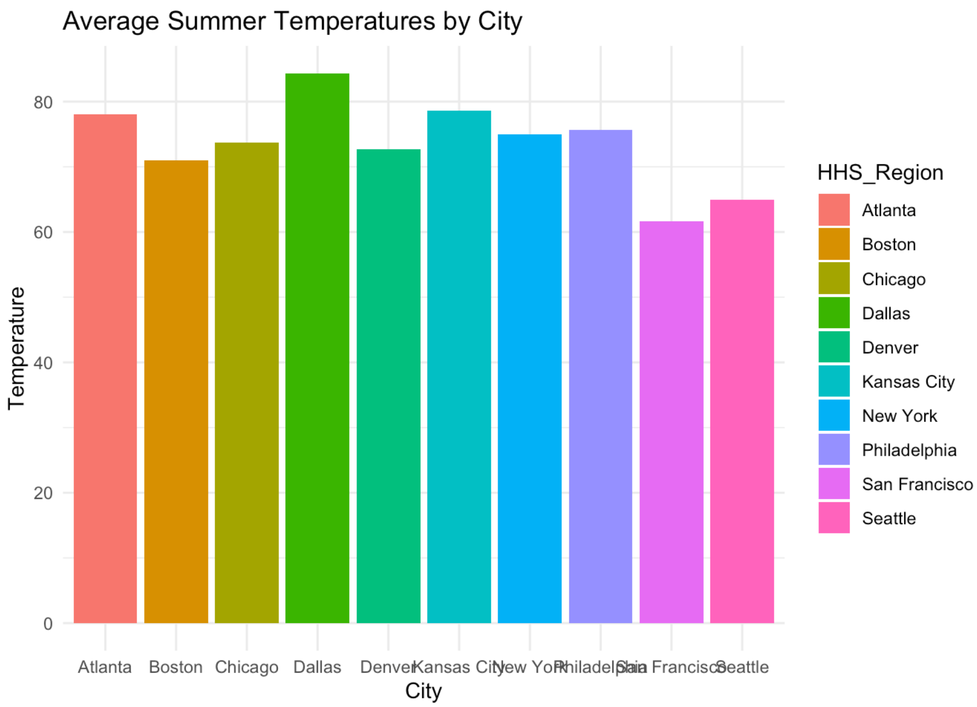

Temperature Trends Across Four Seasons - Atlanta
Temperature
70
60
50
Spring
Summer
Fall
Winter
Season


Temperature Trends Across Four Seasons - Philadelphia
Temperature
70
60
50
40
Spring
Summer
Fall
Winter
Season

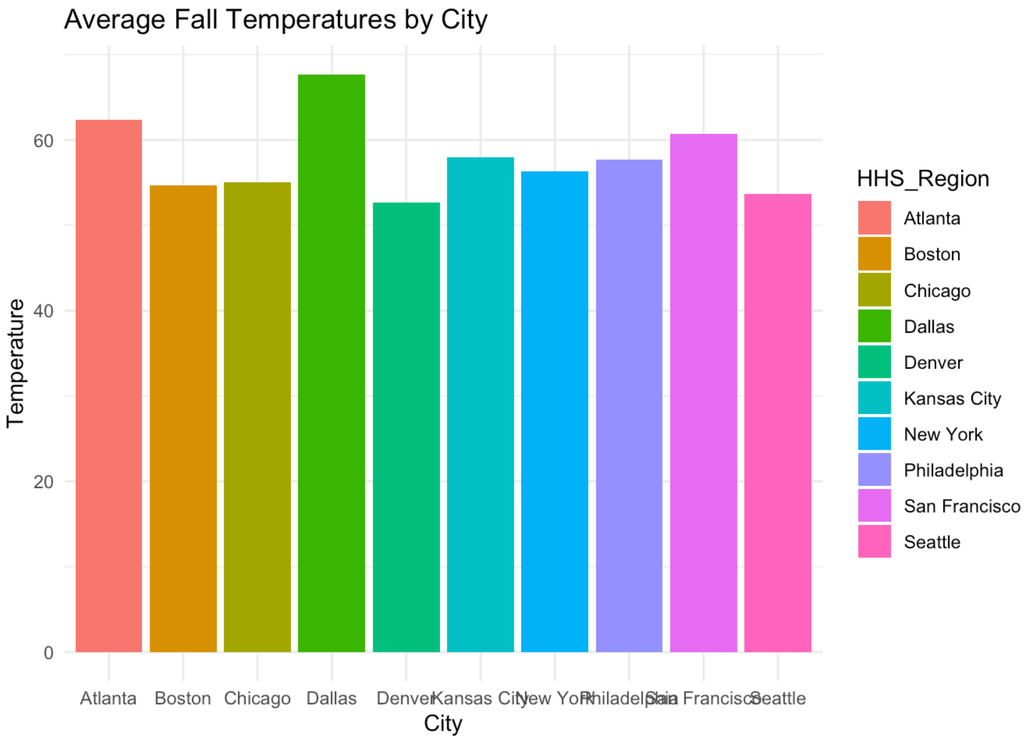

Temperature Trends Across Four Seasons - New York
Temperature
70
60
50
40
Spring
Summer
Fall
Winter
Season


Temperature Trends Across Four Seasons - Seattle
Temperature
65
60
55
50
45
Spring
Summer
Fall
Winter
Season

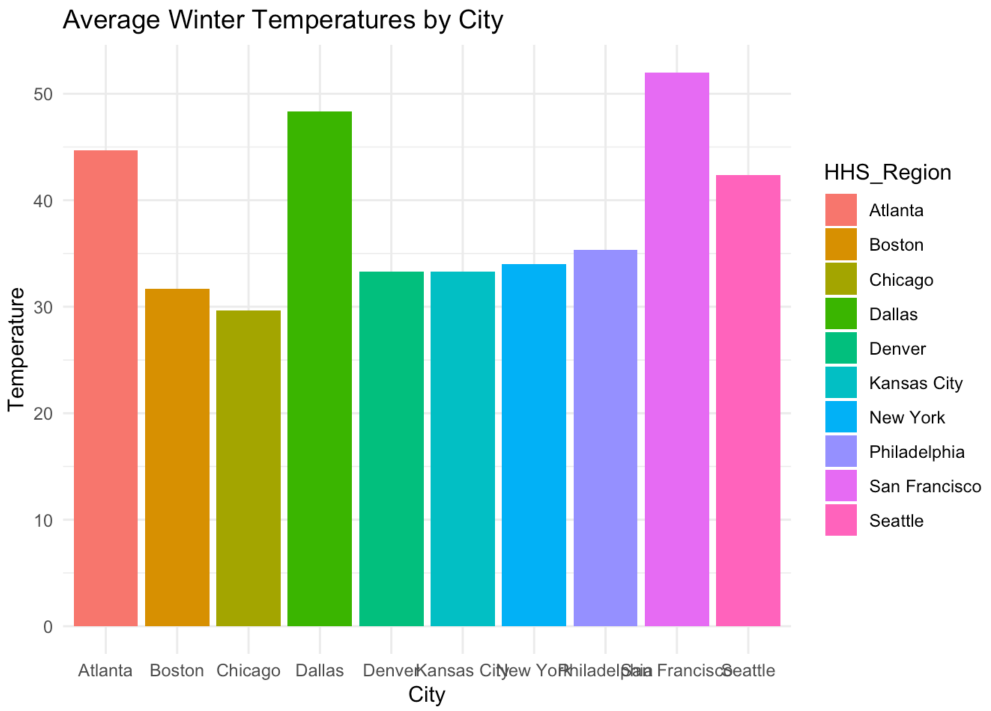

Temperature Trends Across Four Seasons - Boston
Temperature
70
60
50
40
30
Spring
Summer
Fall
Winter
Season



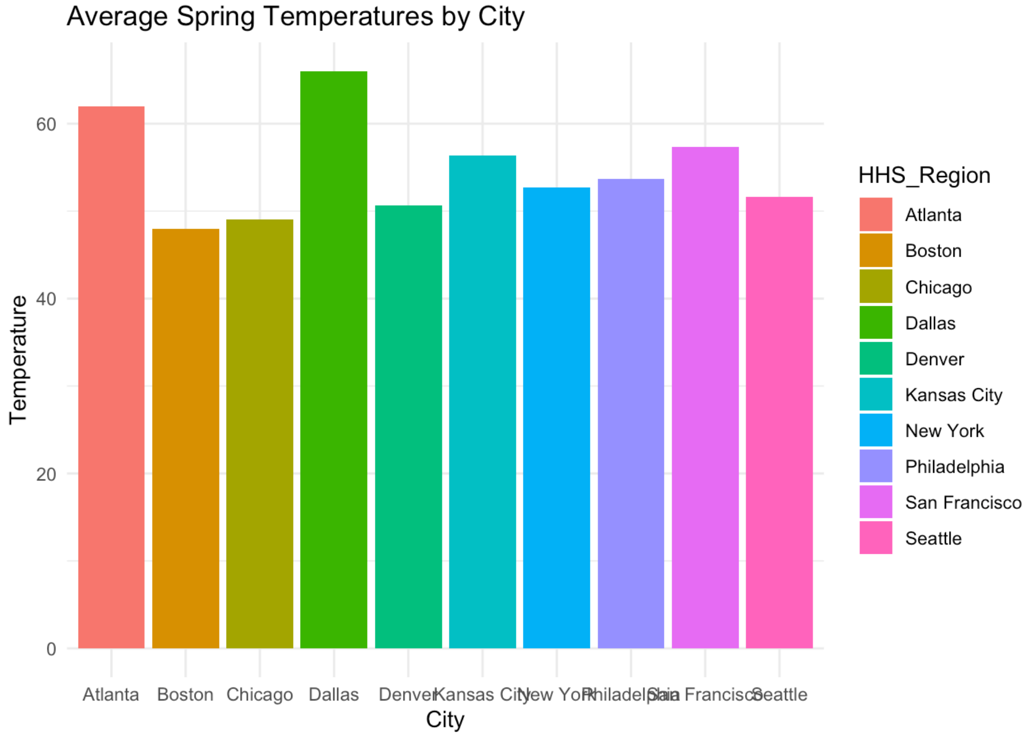
Average Summer Temperatures by City



Average Fall Temperatures by City



Average Winter Temperatures by City



Average Spring Temperatures by City