

Avocado Price and Consumption

Final Project

By

Vanessa Da Veiga



Contents

1. Data Collection
2. Data Wrangling
3. Exploration
4. ANOVAs
5. Logistic Regression
6. Further Exploration



Avocado Consumption

Comparing Prices and Volume of Avocados in three different US Regions

- Boston
- Chicago
- South Carolina

Data Collection Methods

- Downloaded from the Woz-U Github Repo
- CSV File
- Imported using Pandas library
- Original source is from Kaggle
- [census.gov/quickfacts](https://www.census.gov/quickfacts) Tables



Data Wrangling

- Original dataset contained 18249 rows
- Dataset includes 54 regions
- From the years 2015 to 2018
- A subset of 3 regions was used in the analysis
- Sample size for each region was 338

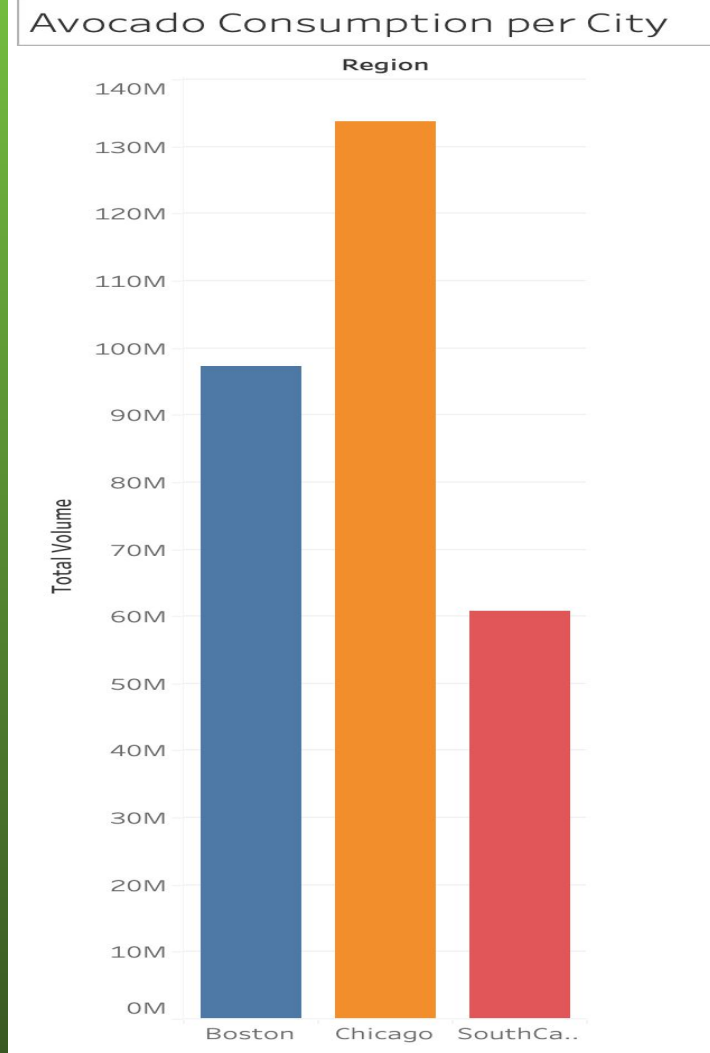
```
df1.region.value_counts()
```

```
Boston          338  
Chicago          338  
SouthCarolina   338  
Name: region, dtype: int64
```

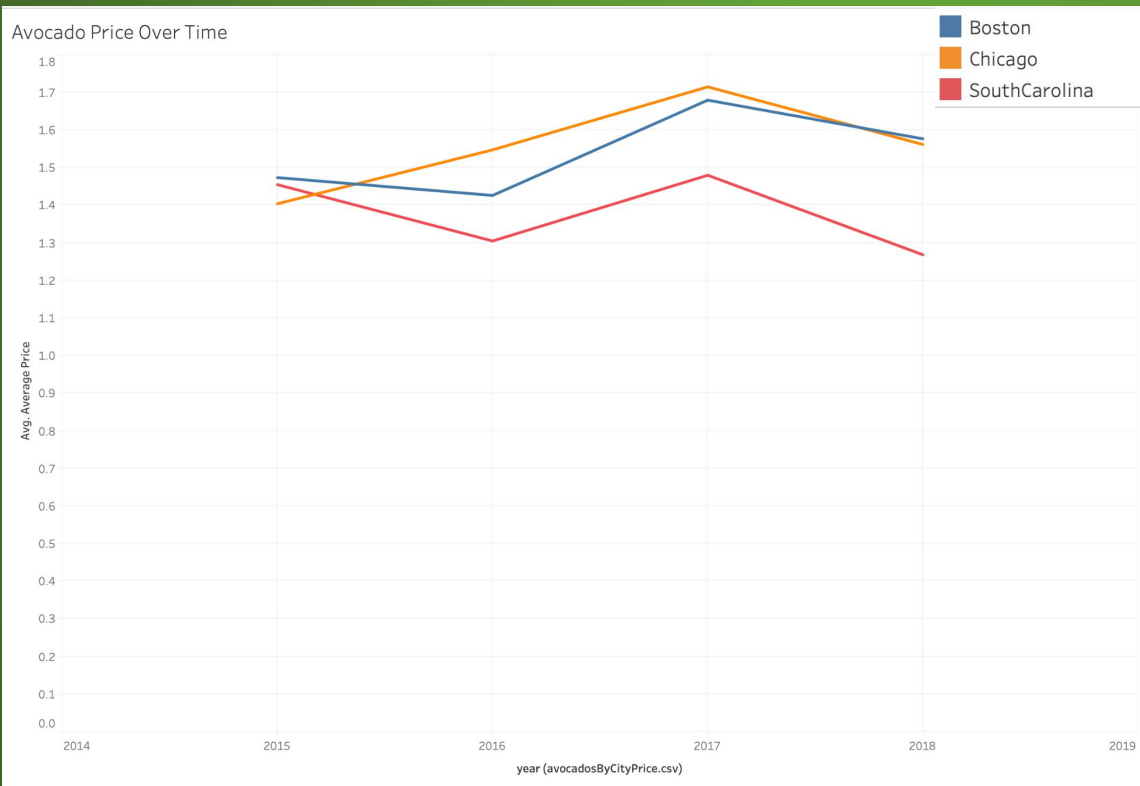
Avocado Consumption per Region

Chicago is in 1st place, Boston in 2nd, and South Carolina in 3rd.

South Carolina having less than half of the consumption measure compared to Chicago.

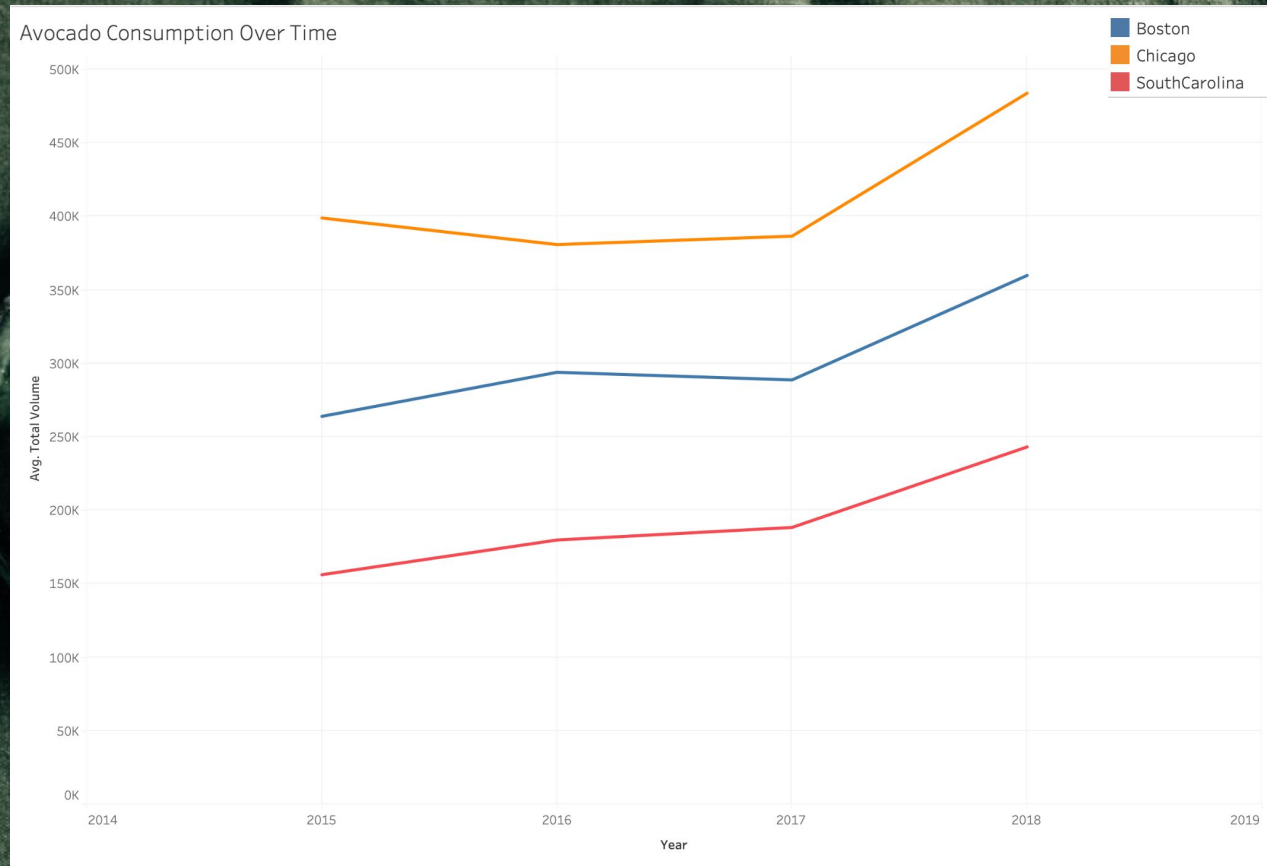


Avocado Prices Over Time



**Average avocado price of all
54 regions for all 4 years
was \$1.41**

Avocado Consumption Over Time

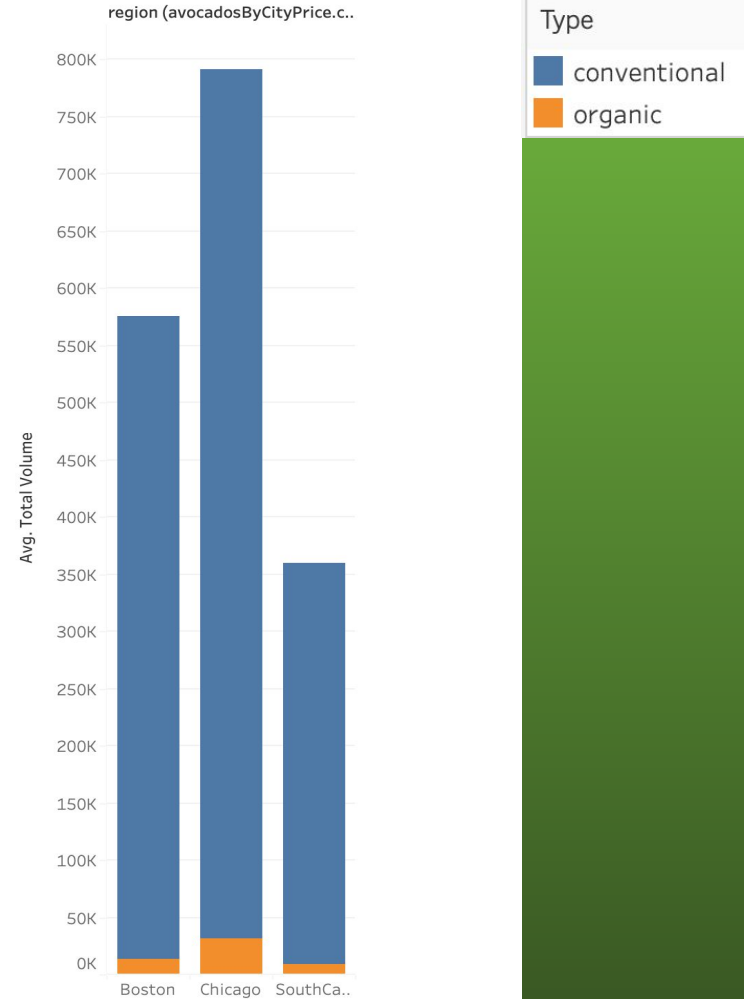


Data Exploration

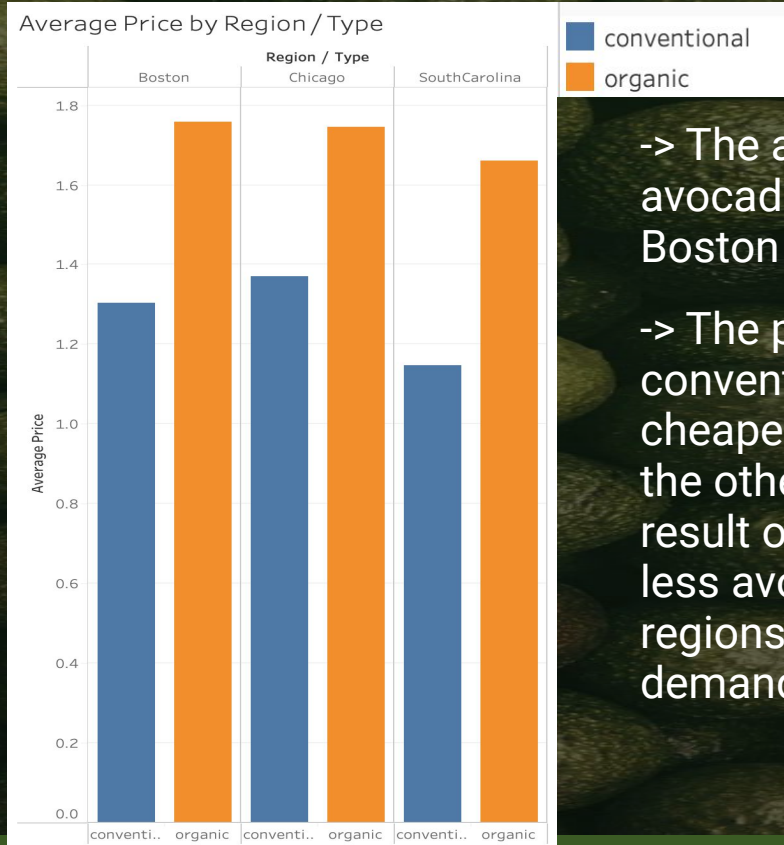
Total Volume by Region

Conventional x Organic

Total Volume by City and Type



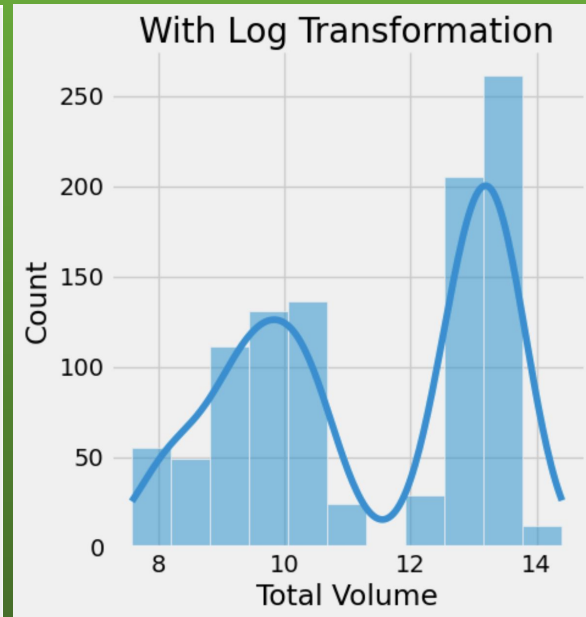
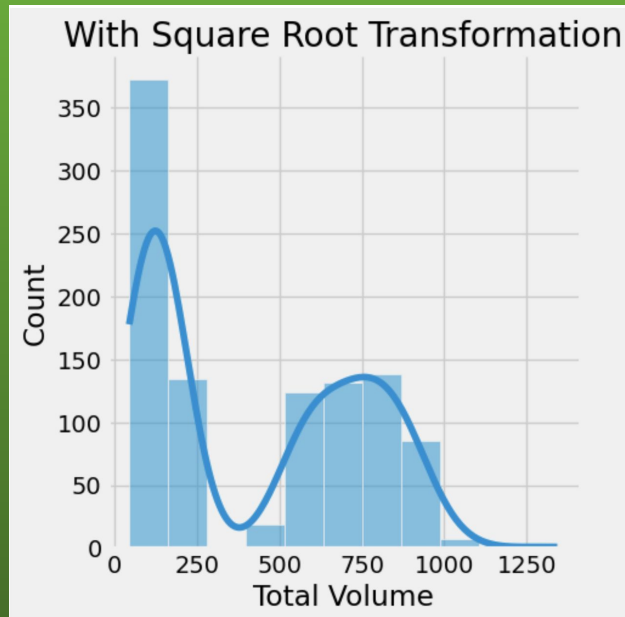
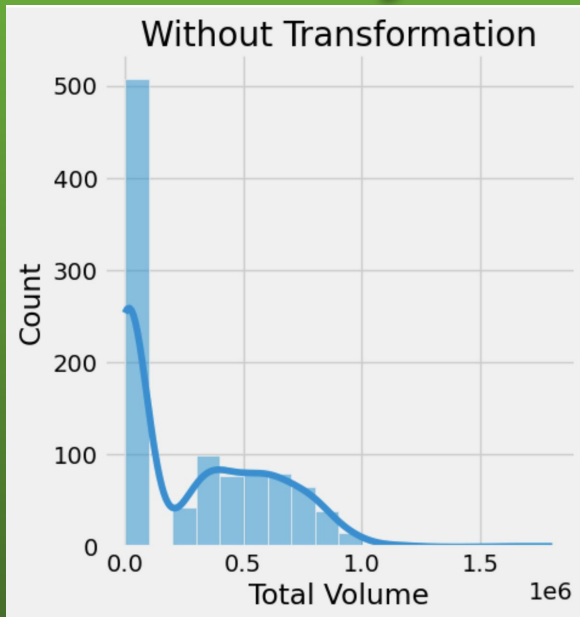
Average Price by Region and Type



-> The average price for organic avocados are very similar between Boston and Chicago.

-> The price for both organic and conventional avocados are slightly cheaper in South Carolina than in the other regions. This could be the result of South Carolina consuming less avocados than the other two regions, therefore having a smaller demand and lower prices.

One-Way ANOVA



Log Transformation normalized the most.

One-Way ANOVA

```
bartlett(df1["Total Volume_log"], df1["regionR"])
```

```
BartlettResult(statistic=681.0293960138611, pvalue=3.990868355374021e-150)
```

The p value is very small and violates this assumption. Will continue.

ANOVA Assumption of Normality - CONT

A Bartlett's test shows that the assumption of homogeneity of variance has been violated. However, we will move forward but be wary of our results.

One-Way ANOVA

- Sample size has at least 20 cases per independent variable.
- No overlap between the groups (regions), and region testing are not related.
- Sphericity is not applicable because it is not a between subjects design.

One-Way ANOVA Comparing Region and Volume Sold

ANOVA Results

- ANOVA Result was significant serving as evidence that there was a difference between the average avocado consumption (volume sold) of the three regions.
- There was also a significance in consumption between each region.

Total Volume	
region	
Boston	287792.854527
Chicago	395569.048846
SouthCarolina	179744.890237

Linear Regression

After doing a linear regression, we have a prediction for the following year (2019) of avocado prices for the three cities we have chosen in the beginning.

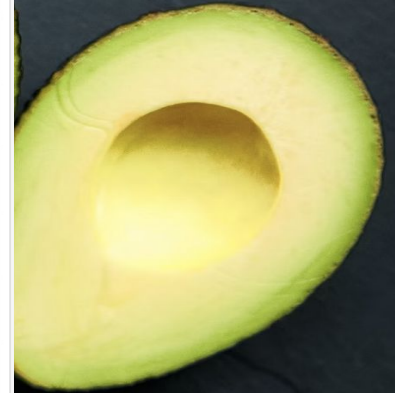
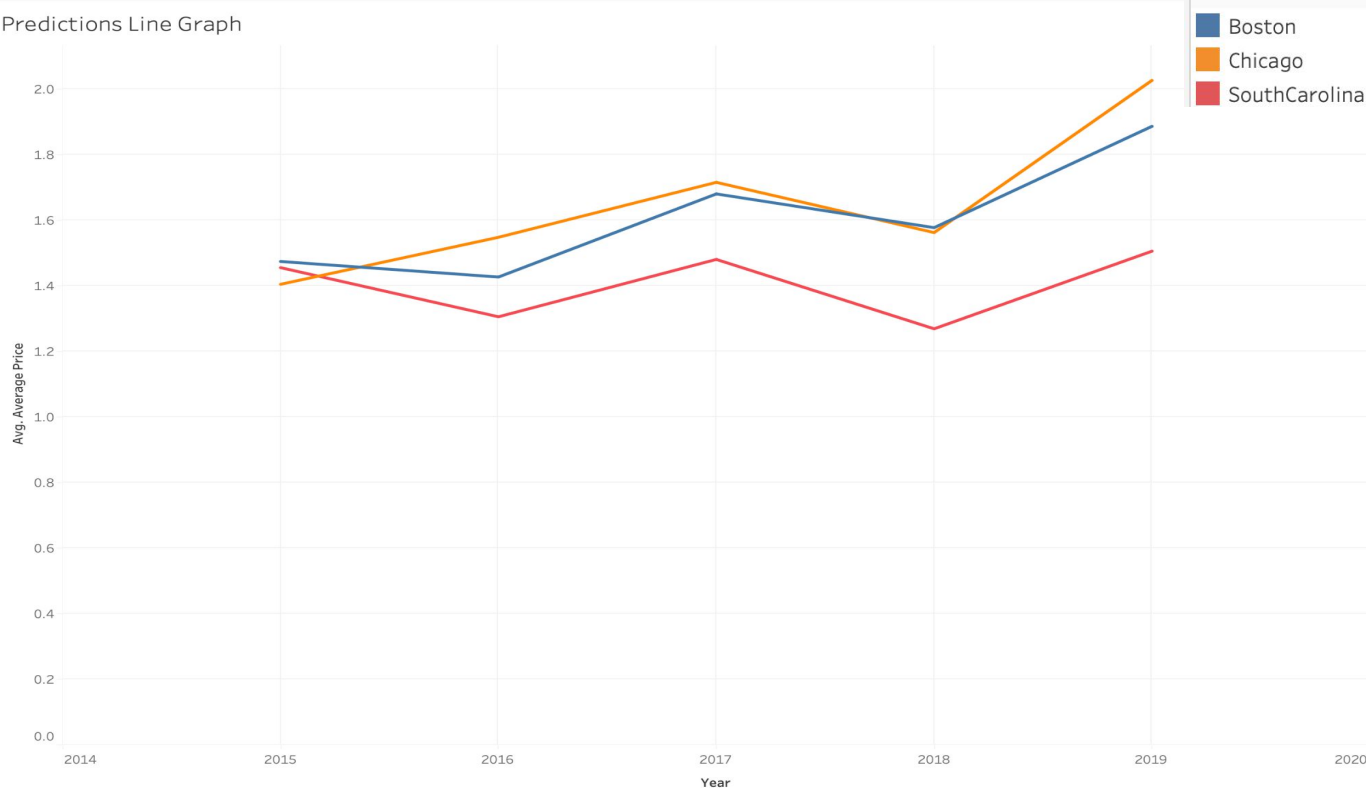
However, the R squared is very much out of range, and that will be kept in mind when looking at these results.

```
Boston - Mean squared error: 0.03, R-squared: -4.74
Chicago - Mean squared error: 0.05, R-squared: -923.81
SouthCarolina - Mean squared error: 0.04, R-squared: -112.83
```

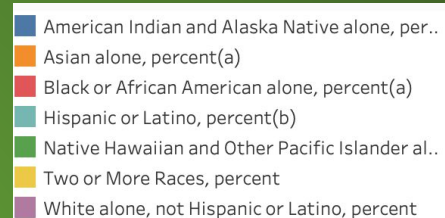
	Region	Year	Predicted_Price
0	Boston	2019	1.885499
1	Chicago	2019	2.025584
2	SouthCarolina	2019	1.504911

Predictions and Original Data Line Graph

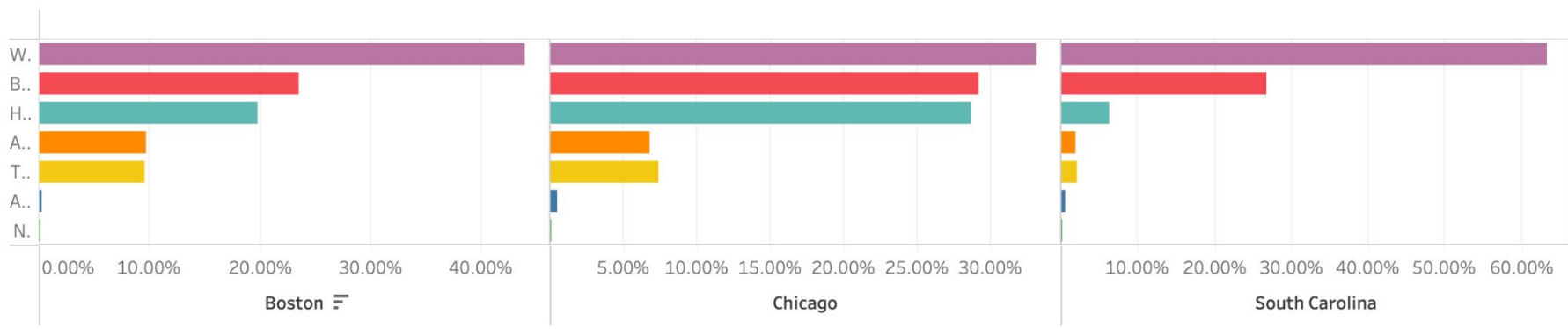
Predictions Line Graph



Race and Ethnicity by Region



Race and Ethnicity by Region





THANK
YOU

