Minnesota High School Data Analytics (2018-2019)

Content

Business Case: Analysis of Minnesota High Schools Standard test score performance during the academic year 2018-2019

Data Acquisition: Dataset was obtained in a CSV format

Data Preparation: Using Azure ML, cleaned all missing data

Data Visualization: Descriptive Statistics using Tableau

Data Visualization - Tableau

Tableau is an interactive data visualization tool used for Exploratory Data Analysis (EDA), where charts are plotted using independent variables (dimensions – qualitative values) against measures (quantitative values) and dependent variables (readmit30) to get insights and understand their data. Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods

Tableau is quick, simple, user-friendly, intuitive, can handle lot of data, provide statistical calculations on datasets

EDA:

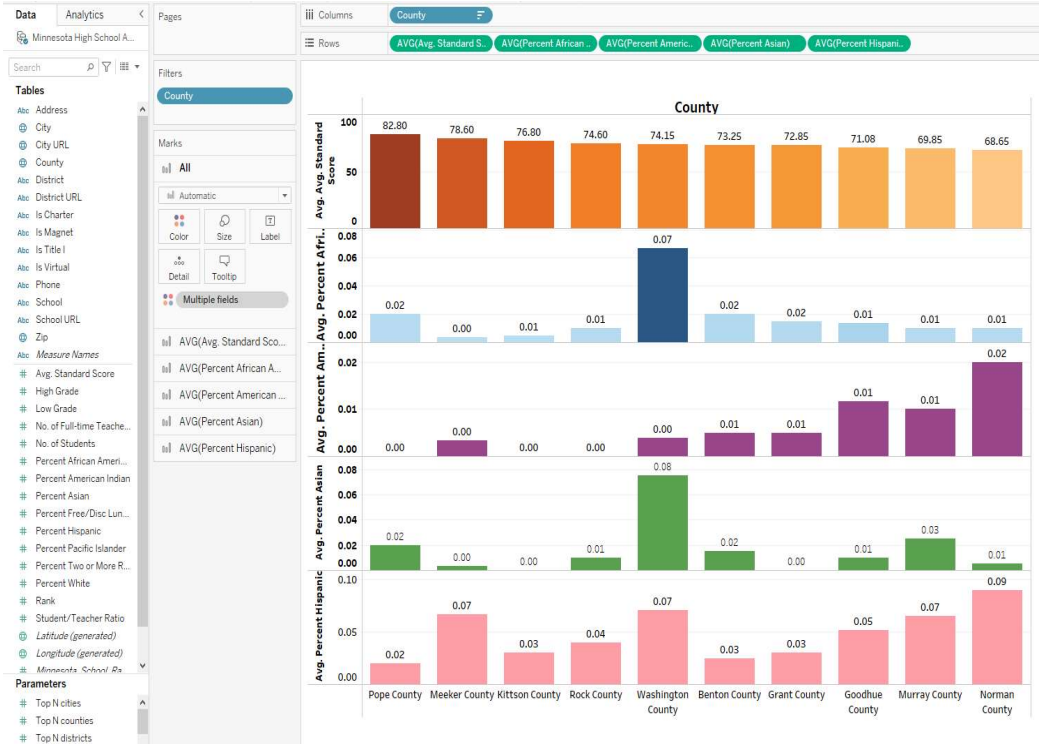
- ❑ Get a better understanding of data that may not be analyzed by standard data science algorithms.
- ❑ Understanding data patterns that may be skipped by typical machine learning algorithms.
- ❑ Drawing charts and graphs for better understanding from different angles and projects the results as charts and graphs.
- ❑ To get a better understanding of the problem statement, with graphs and charts.
- ❑ To find the hidden trends and relationship between variables.
- ❑ Assess and validate your assumptions on the variables, whether the variables help answer business problem or not.
- ❑ Screen for noise variables, missing data, outliers, etc. Find which variables need imputation, preprocessing

Tableau

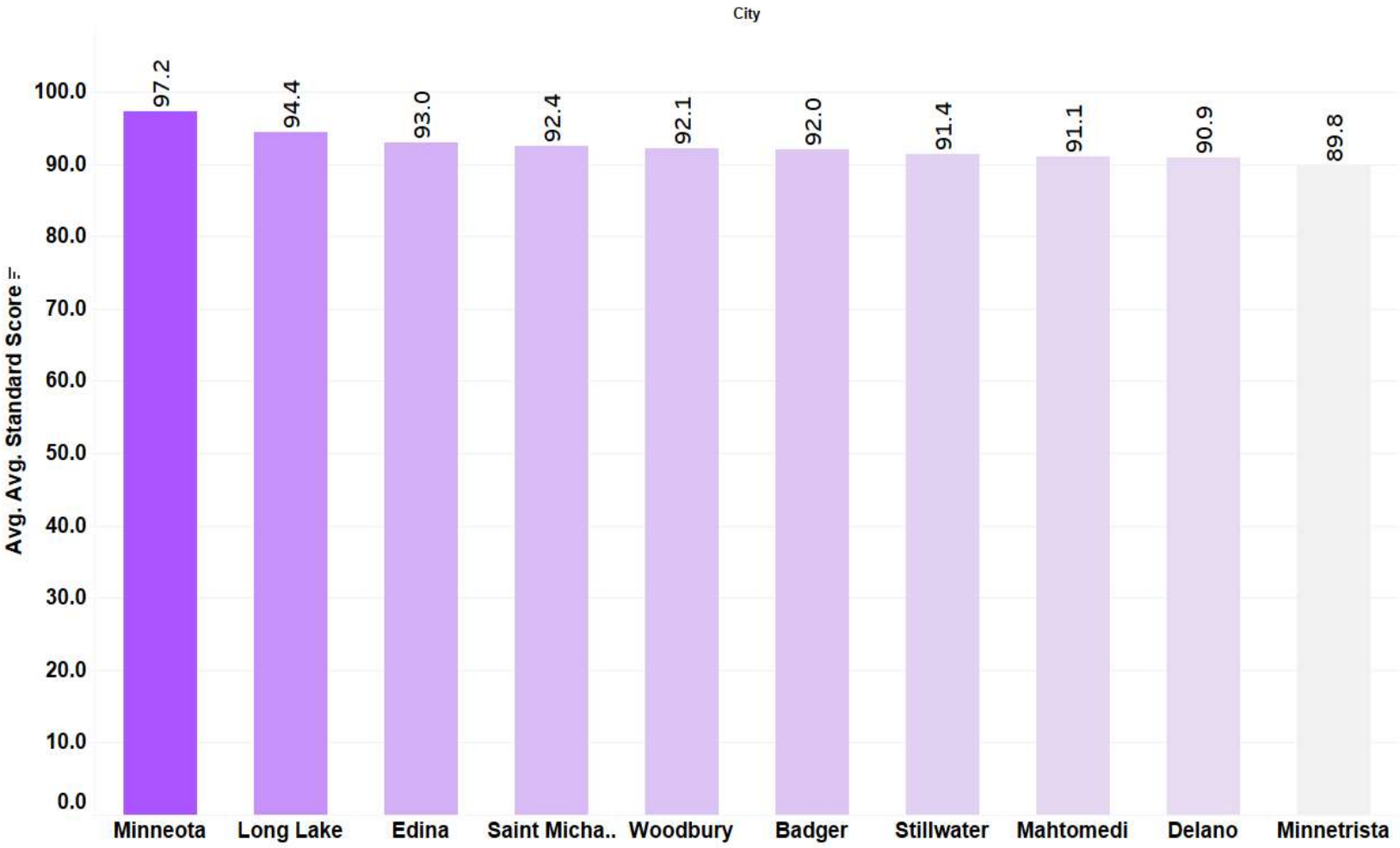
Chart Views

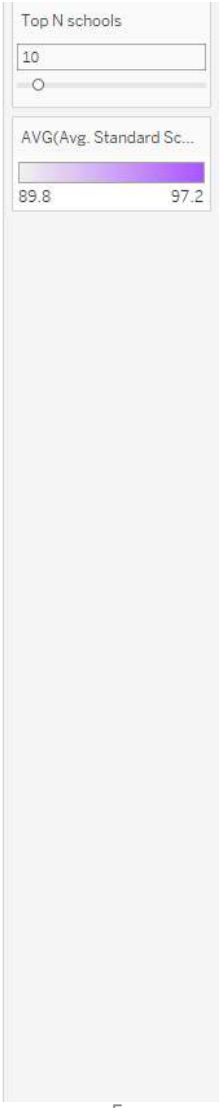
1. Text tables
2. Heat maps
3. Highlight tables
4. Symbol maps
5. Maps
6. Pie charts
7. Horizontal bars
8. Stacked bars
9. Side-by-side bars
10. Tree maps
11. Circle views
12. Side-by-side circles
13. Lines (continuous)
14. Lines (discrete)
15. Dual lines
16. Area charts (continuous)
17. Area charts (discrete)
18. Dual combination
19. Scatter plots
20. Histogram
21. Box and whisker plots
22. Gantt
23. Bullet graphs
24. Packed bubbles

Data Pane, Marks card and Worksheet

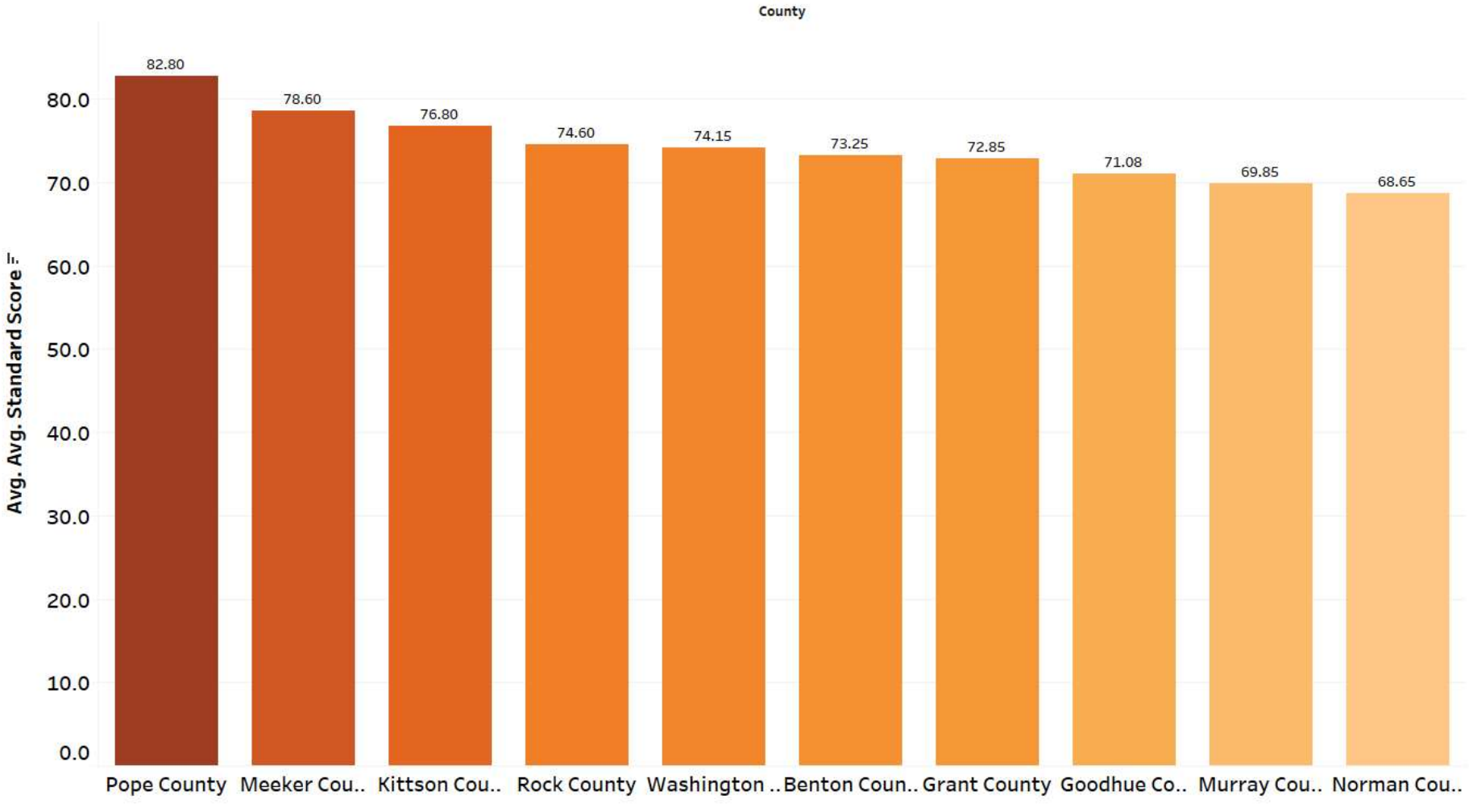


Top 10 cities with highest Avg. Standard score

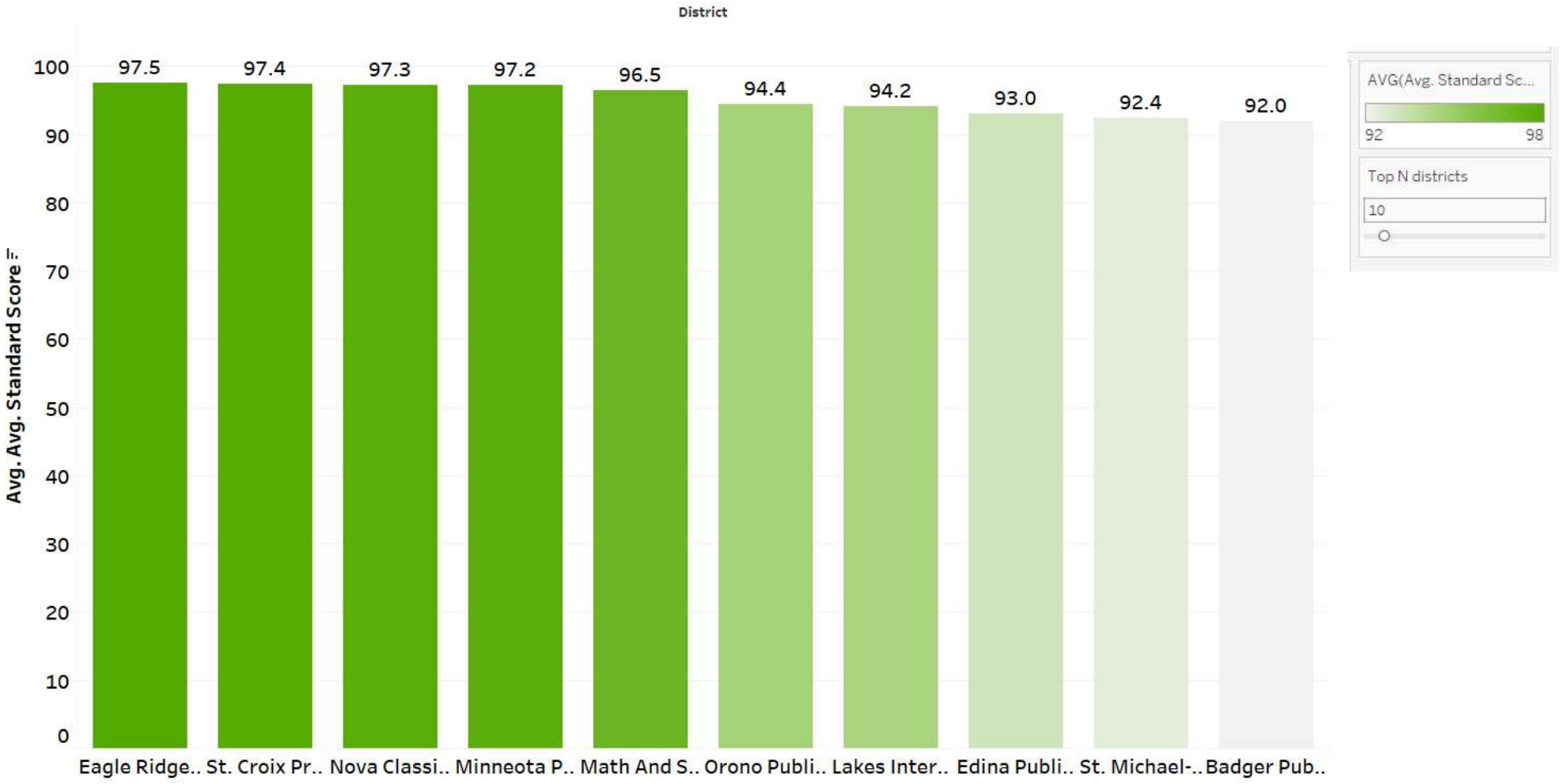




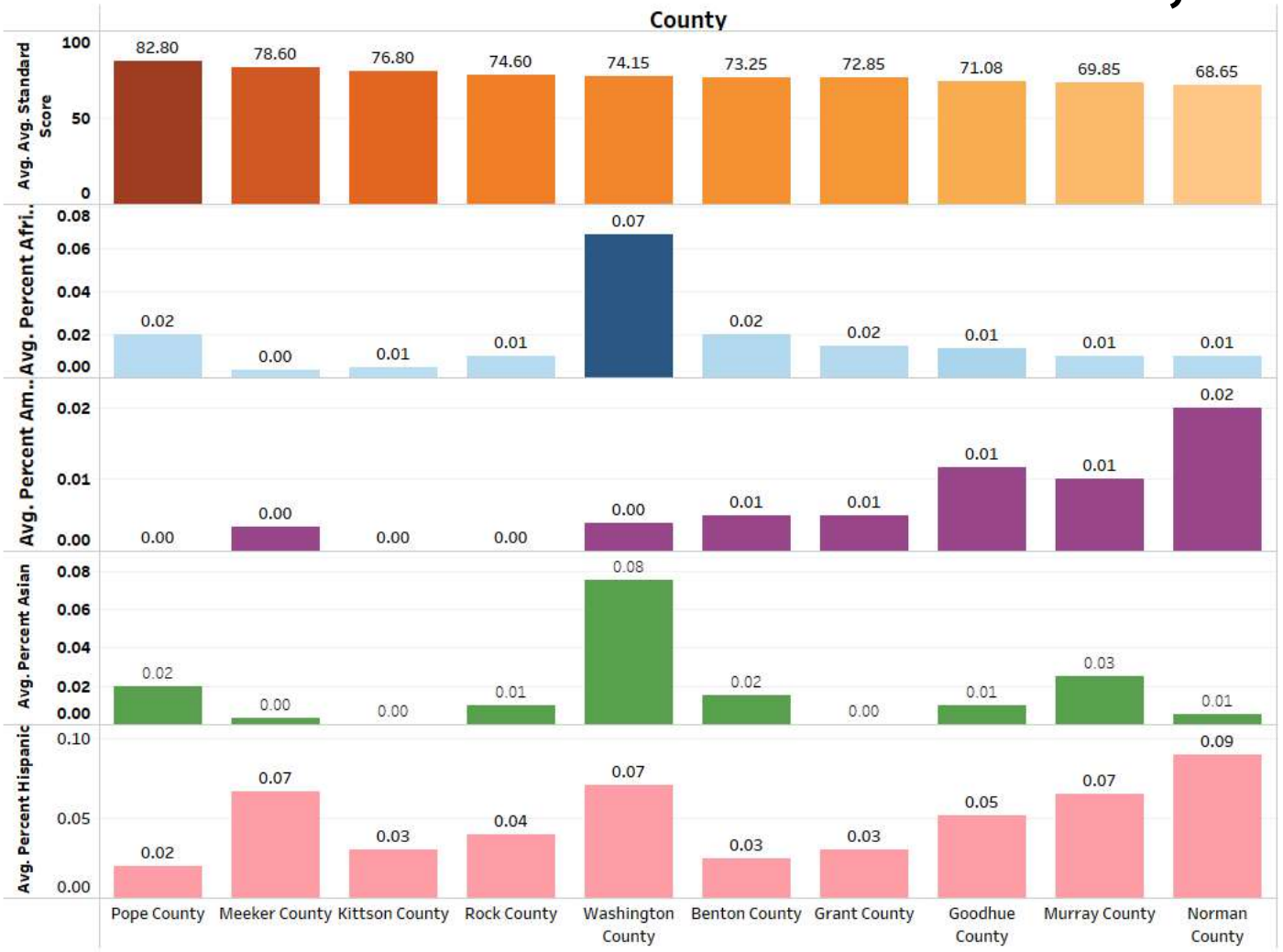
Top 10 counties with highest Avg. Standard score



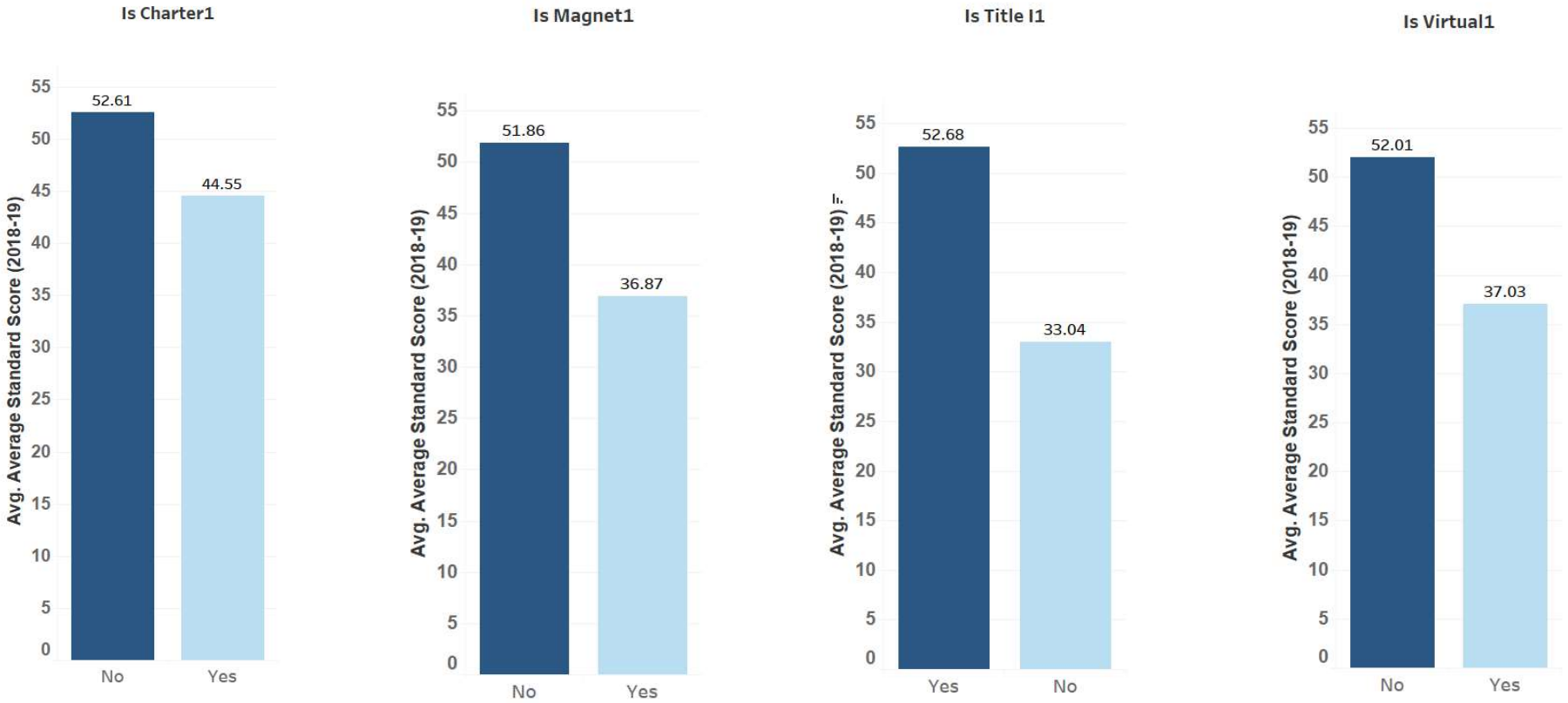
Top 10 districts with highest Avg. Standard score



Top 10 counties Vs Standard score, Race

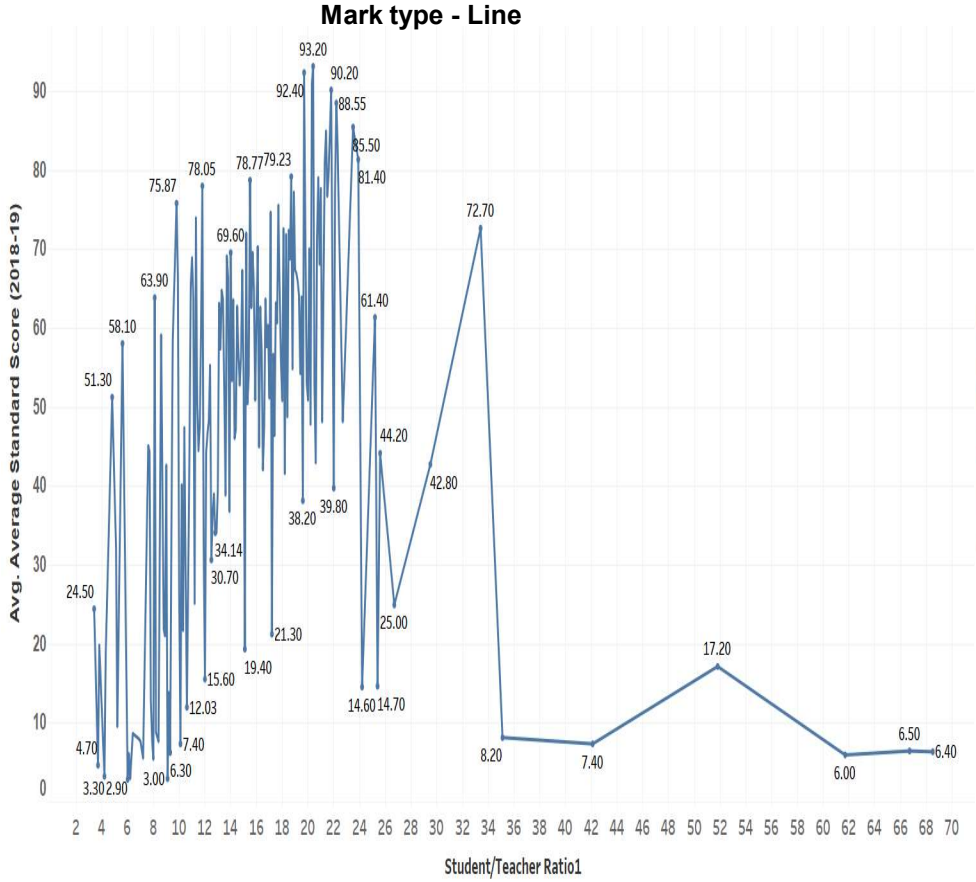


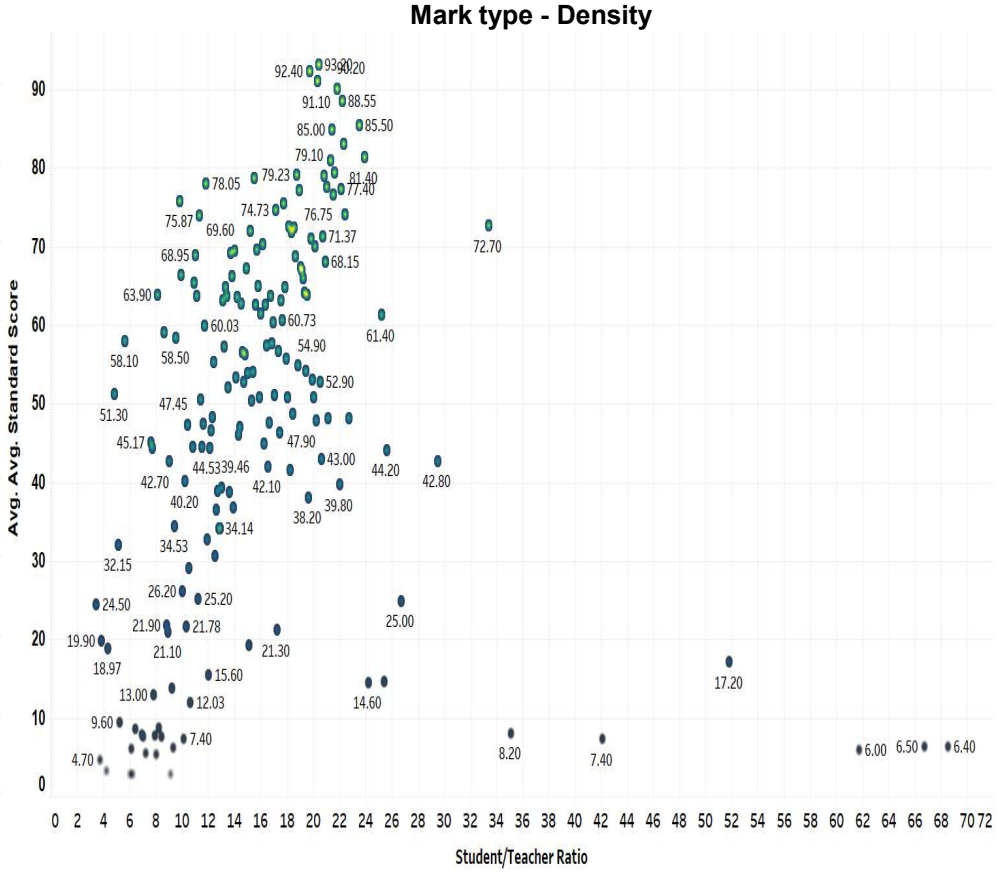
Charter, Magnet, Title and Virtual schools



Title, Charter, Magnet and Virtual schools .

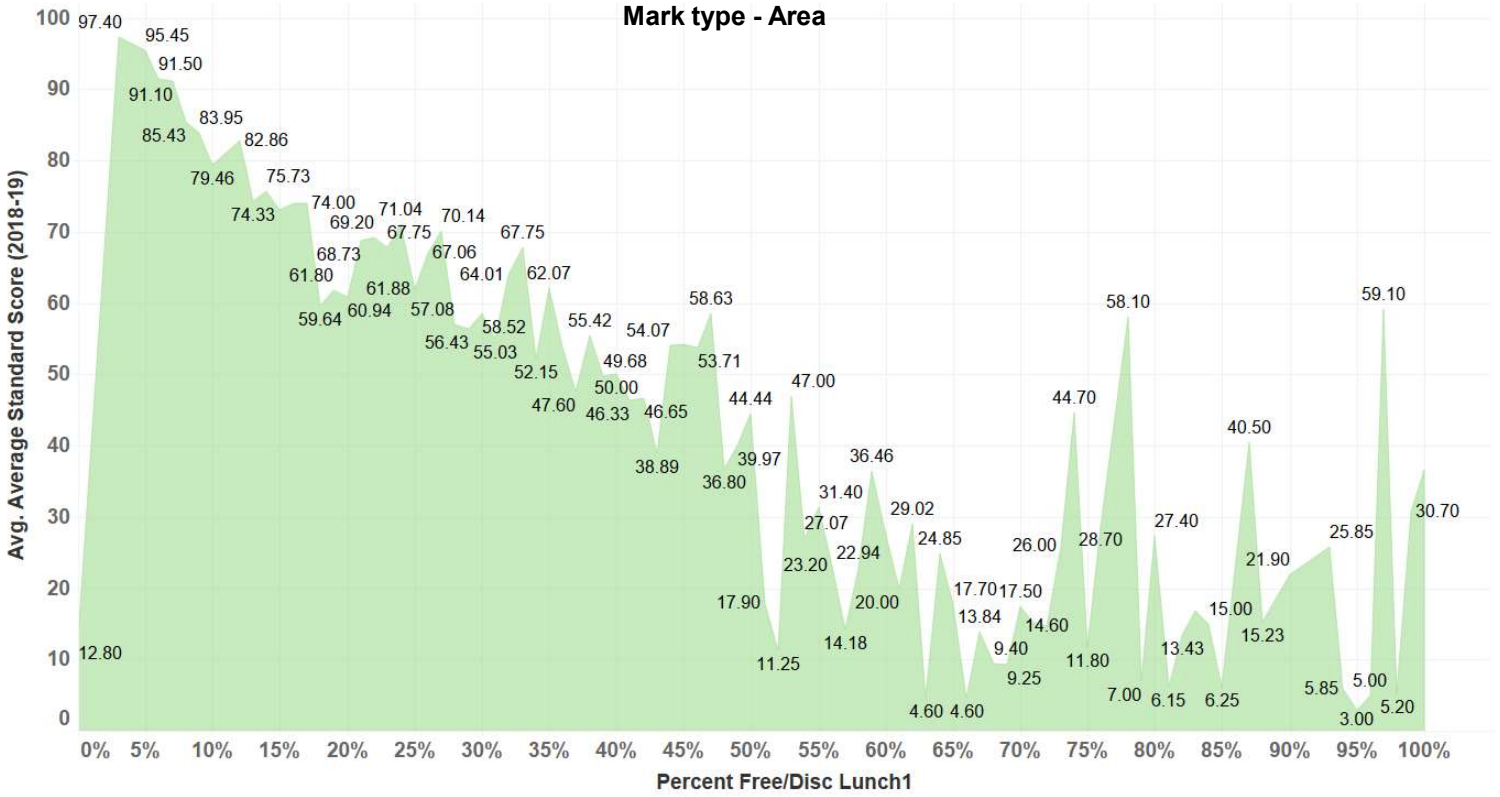
Student/ Teacher ratio Vs Standard score





From the chart, it is inconclusive if Student/Teacher ratio affects test scores.

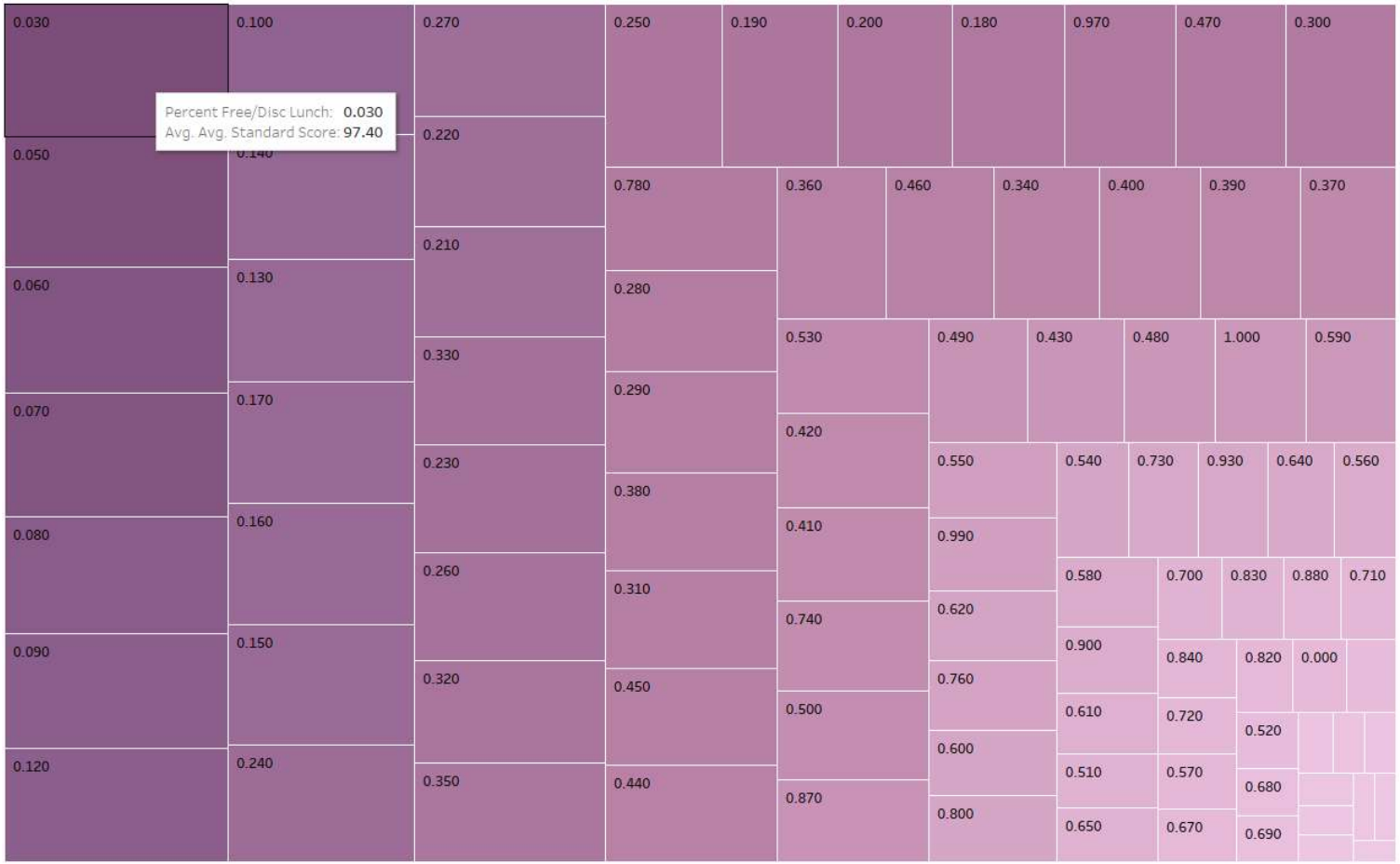
Percent Free/Disc Lunch Vs Std. score



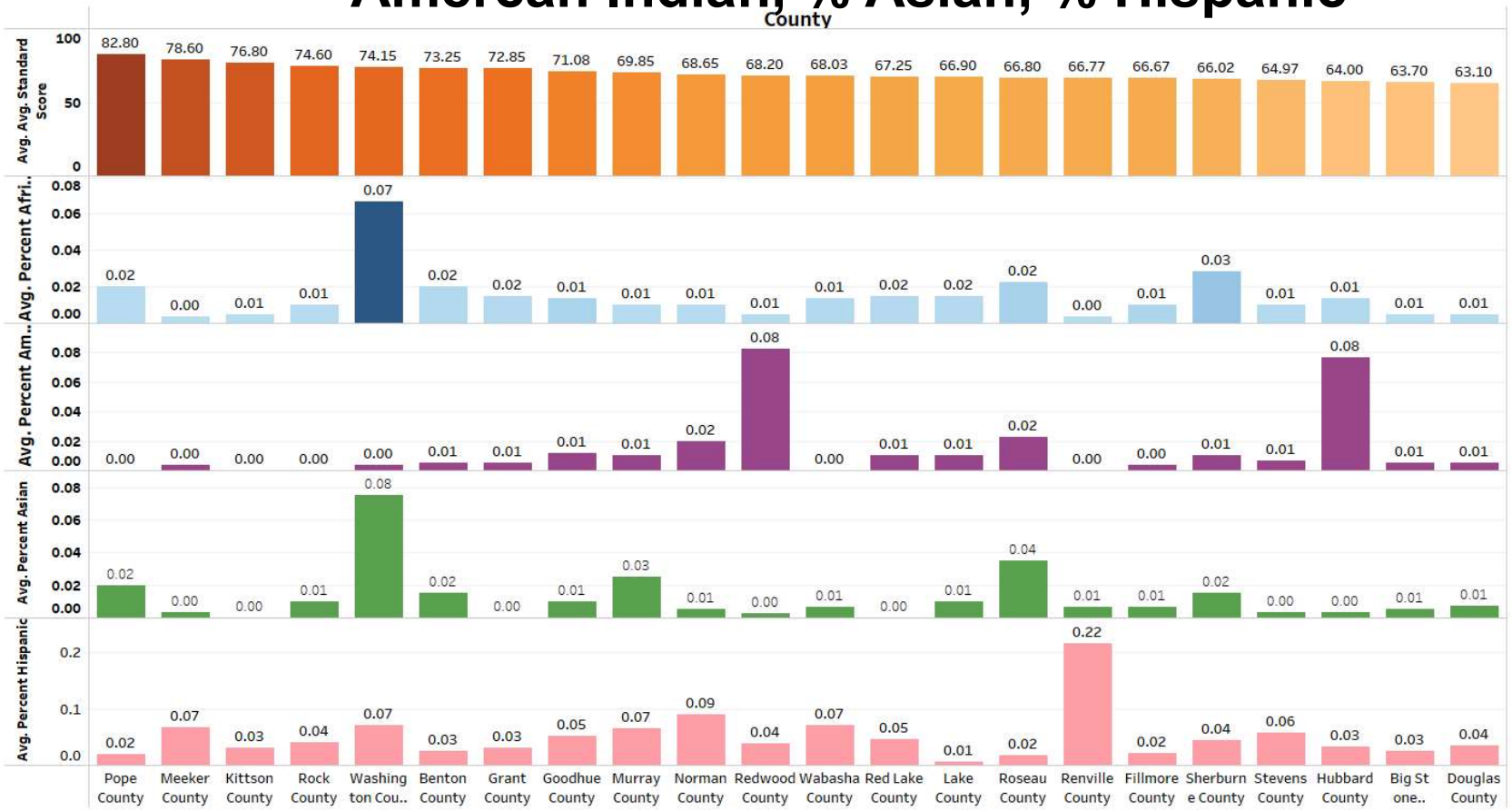
The chart shows, test scores in schools decreases with increase in percent free/ discounted lunches up to 50th percentile. After 50th percentile, its inconsistent.

Percent Free/Disc Lunch Vs Std. score

Mark type - Treemap



County Vs Avg, Standard score, %African American, % Amercan Indian, % Asian, % Hispanic



Data Preparation

Total number of features = 29
Total number of records = 491

No duplicate values found

Missing values
Full-time Teachers (4 records)
Student/ Teacher Ratio (4 records)

Unique Value
High grade has only one unique value (12), all other features have more than one unique value.

Exploratory Data Analysis

The Features (i.e., Variables) are segregated into three categories:

Dependent Variable (Y): Variable that is being measured in the experiment. It changes as a result of the changes to the independent variables. Y values to predict:

Y : Rank and **Standard score**

Noise: Variable that does not affect the dependent variable.

Independent Variable or Predictor Variable (X): Variable whose change isn't affected by any other variable in the experiment.

Independent variable is the cause, and dependent variable is the effect.

Exploratory Data Analysis - continued

Noise Variables

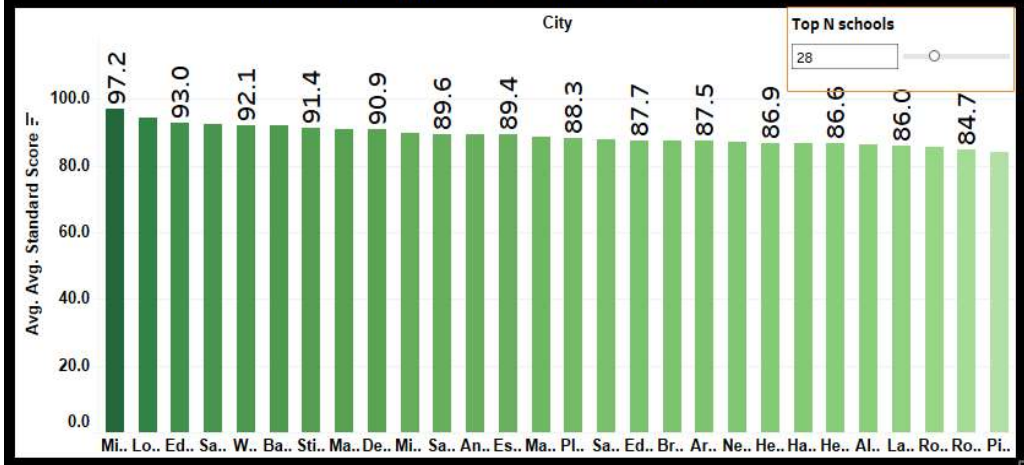
- ❑ School
- ❑ School URL
- ❑ District URL
- ❑ Address
- ❑ City URL
- ❑ Phone
- ❑ High grade

Independent Variables

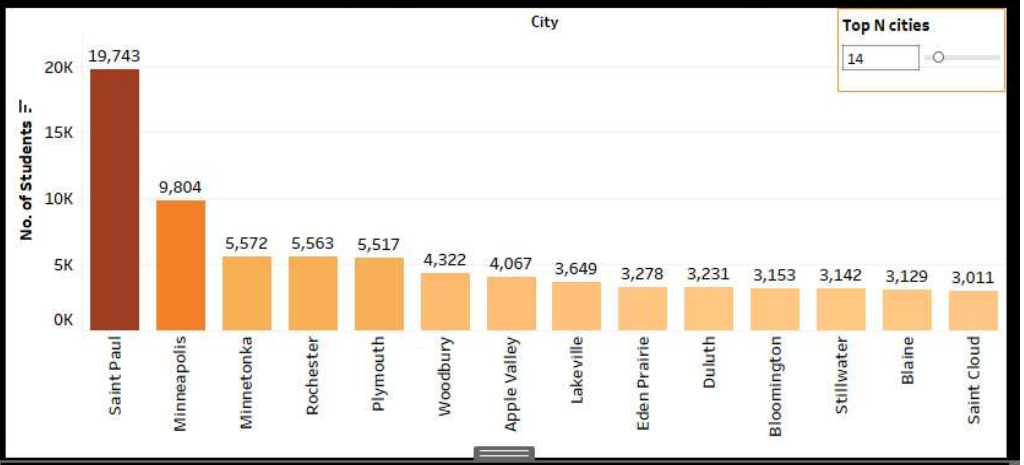
- ❑ District
- ❑ City
- ❑ Zip
- ❑ County
- ❑ Low grade
- ❑ Is Title
- ❑ Is Charter
- ❑ Is Magnet
- ❑ Is Virtual
- ❑ No. of students
- ❑ No. of full-time teachers
- ❑ Student/Teacher ratio
- ❑ Percent free/disc lunch
- ❑ Percent African American
- ❑ Percent American Indian
- ❑ Percent Asian
- ❑ Percent Hispanic
- ❑ Percent Pacific Islander
- ❑ Percent two or more races
- ❑ Percent White

Minnesota High School Analysis (2018-2019) - 1

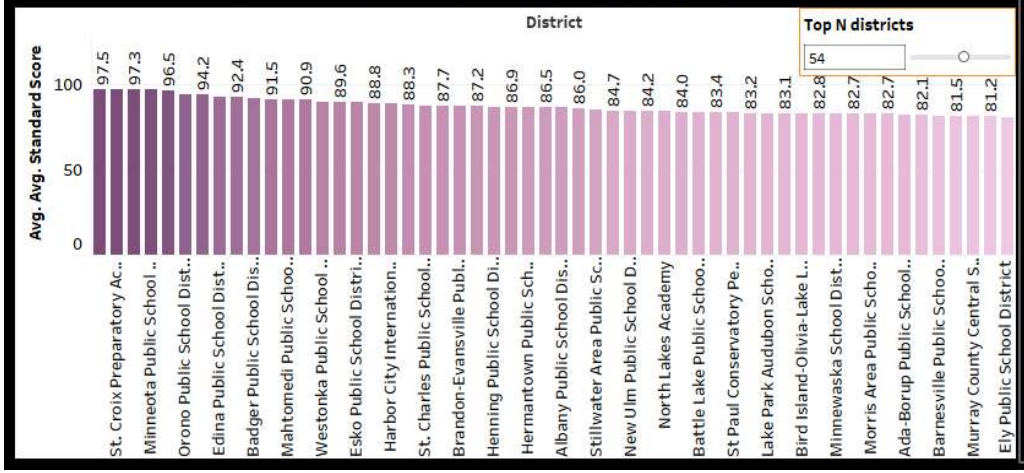
Top 10 cities with highest Avg. Standard score



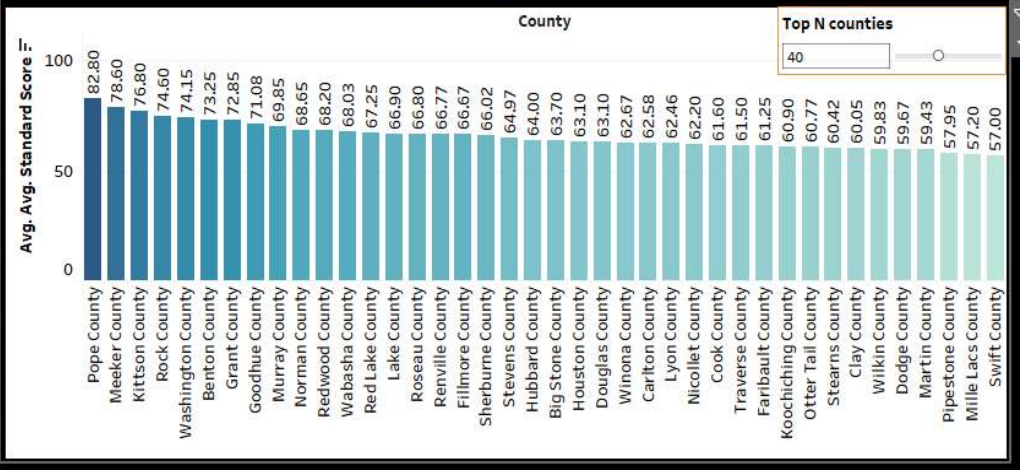
Top 10 cities with highest number of students



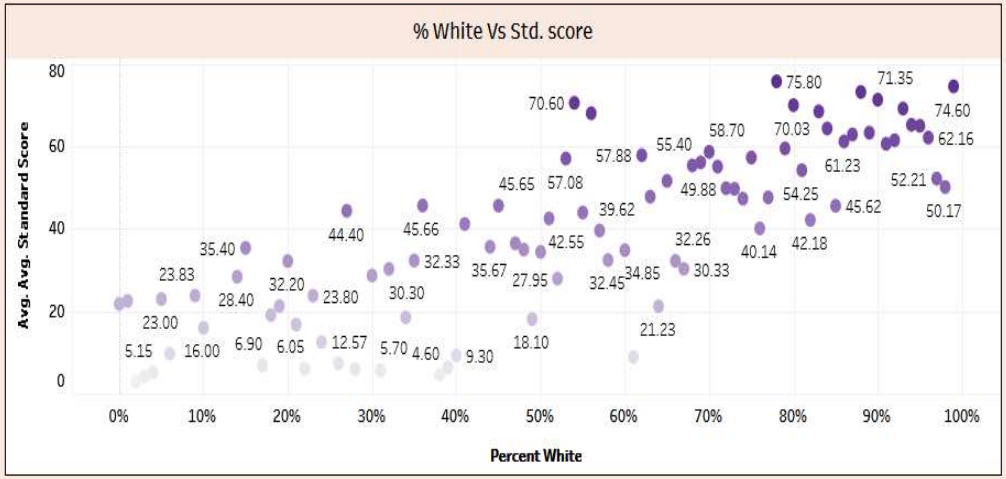
Top 10 districts with highest Avg. Standard score

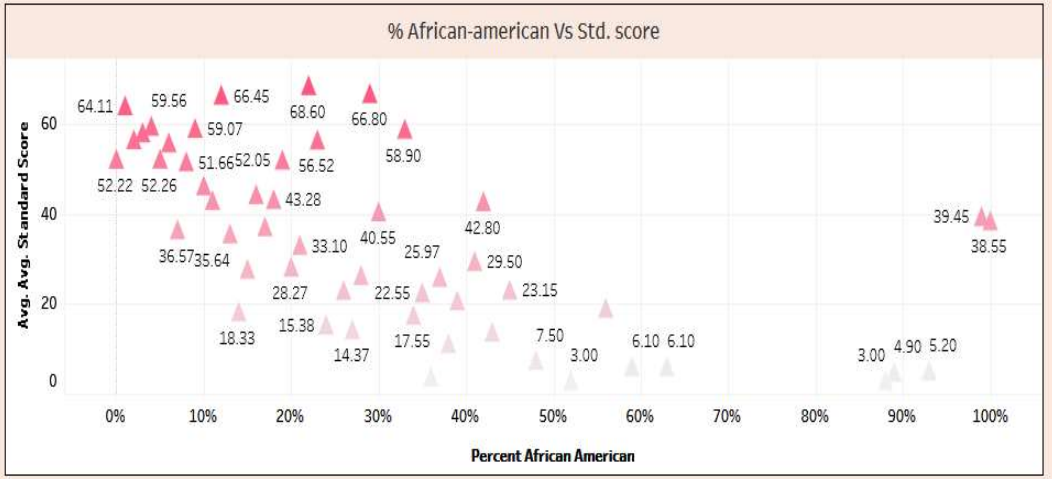


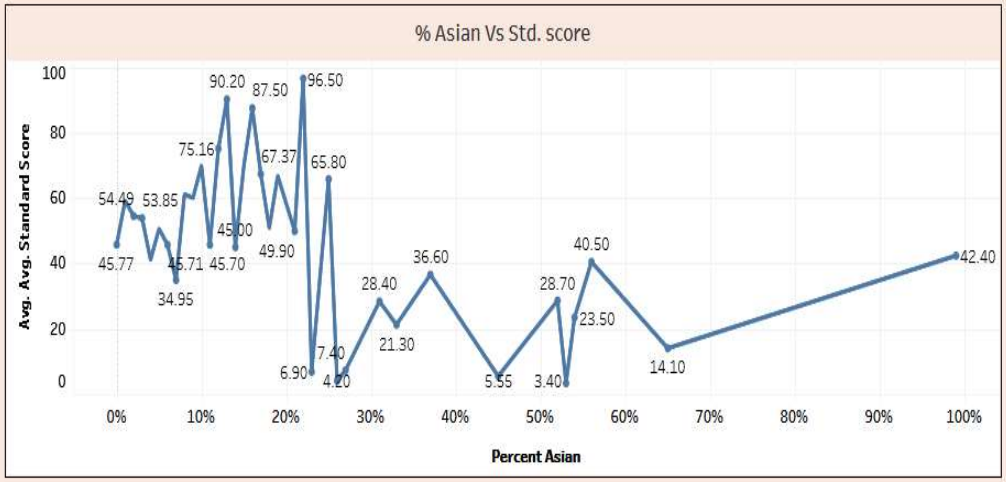
Top 10 counties with highest Avg. Standard score

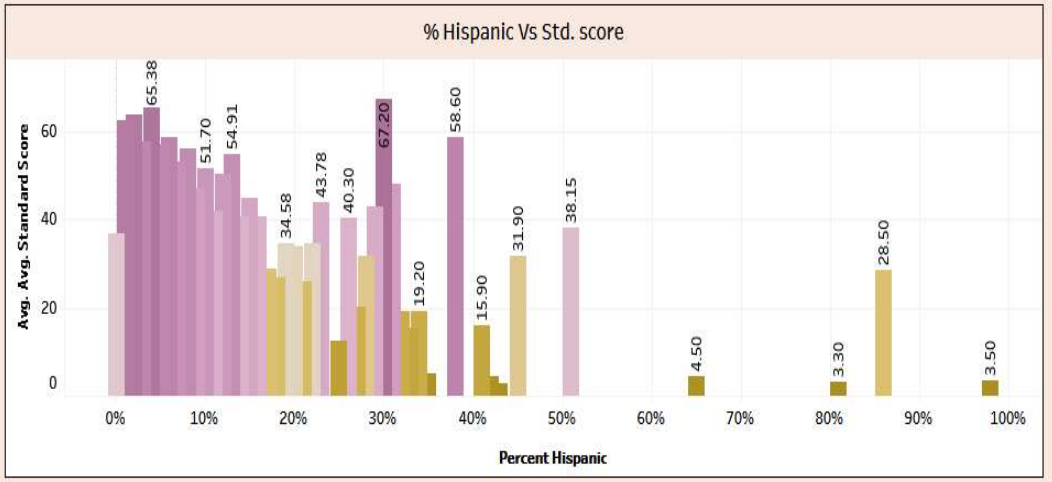


Minnesota High School Analysis (2018-2019) - 2









High School Chi-Squared Test Analysis

Chi-squared test is a statistical method that measures how close expected values are to actual results.

Top 5 impacting features on students test score: **District, City, County, % Free/Disc. Lunch, Number of students**

Bottom 5 impacting features on students test score: **: Percent Pacific Islander, Is Magnet, Is Virtual, Is Title, Is Charter**

The **10 top impacting features** are listed below:

Independent variable	Chi-squared test value
District	3458.642972
City	3144.003962
County	829.108435
Percent Free/Disc. lunch	454.738338
No. of students	256.027904
Percent White	244.519652
No. of full-time teachers	243.003759
Student/Teacher ratio	185.366377
Percent African-American	172.601077
Percent Hispanic	171.361604

High School Linear Correlation Tests Analysis

The correlation coefficient **r** measures the strength and direction of a linear relationship between two variables.

r is always between +1 (Strong positive) and –1 (Strong negative).

Strong correlation: r > 0.7, Moderate correlation: 0.6 to 0.4 , Weak correlation: r < 0.4

Top 3 features that have strong linear relationship with **AVERAGE STD TEST SCORE** of schools: Percent Free/Disc Lunch, Percent White, Percent Hispanic

All other correlations are either weak or moderate.

Independent Variable	R (Independent variable, Average Standard Score)
Percent Free/Disc. lunch	-0.666592
Percent White	0.536807
Percent Hispanic	-0.370136
No. of students	0.367106
Percent African-American	0.353055
No. of full-time teachers	0.335513
Percent 2 or more races	-0.256960
Percent American-Indian	-0.213414
Percent Asian	-0.0991049
Percent Pacific-Islander	-0.065140
Student/Teacher ratio	-0.0348331

Questions?