Healthcare

Description

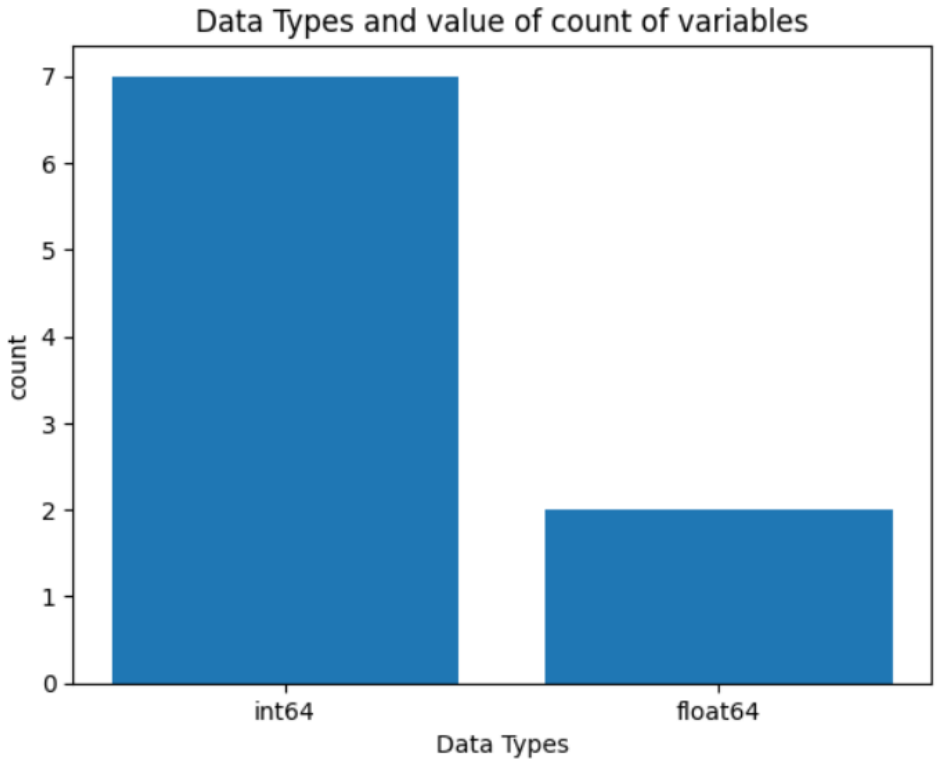
NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.

The objective is to build a model to accurately predict whether the patients in the dataset have diabetes or not. The dataset consists of several medical predictor variables and one target variable (Outcome). The variables include;

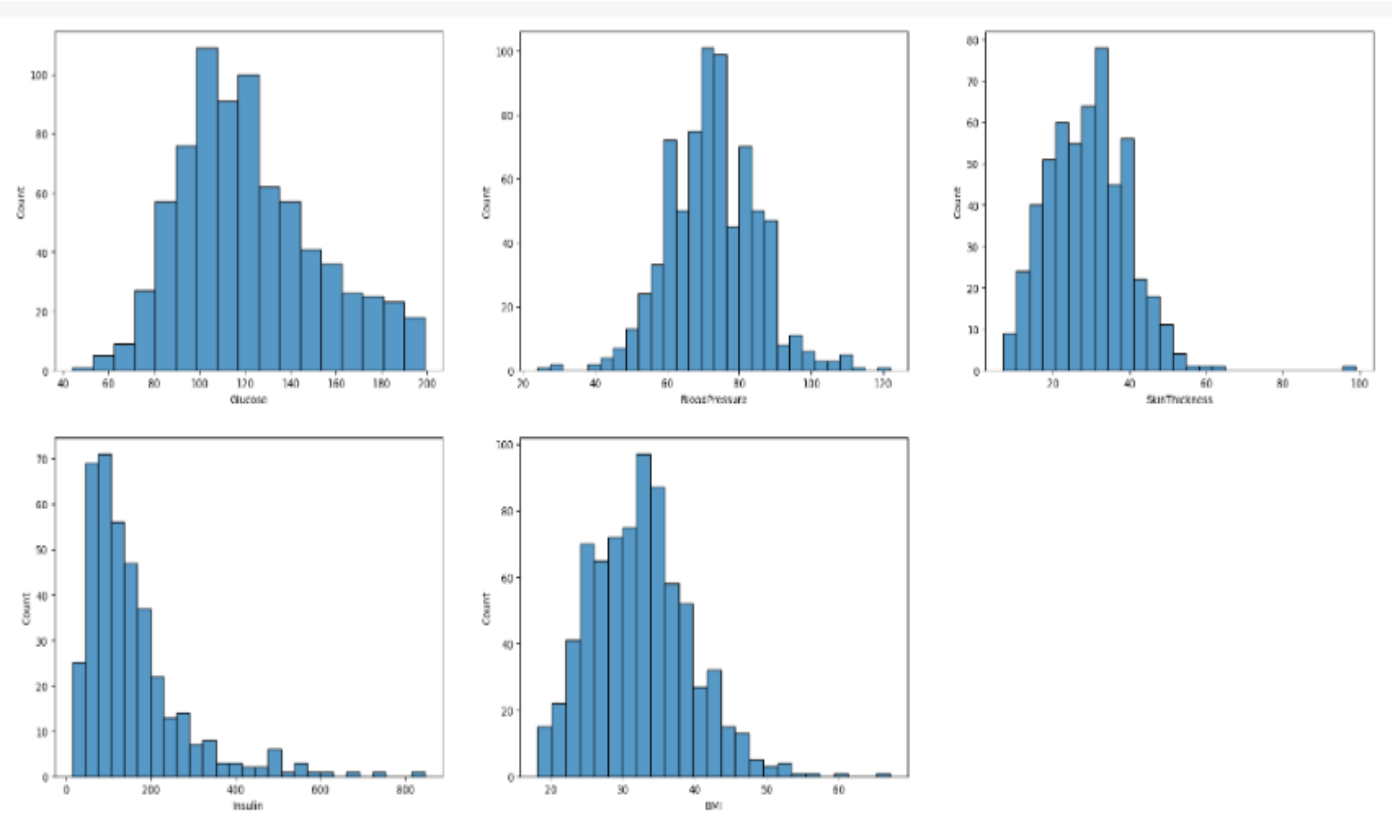
Pregnancies: Number of times Pregnant
Glucose: Plasma glucose concentration in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skinfold thickness (mm)
Insulin: Two-hour serum insulin
BMI: Body Mass Index
Diabetes Pedigree Function: Diabetes Pedigree Function
Age: Age in Years
Outcome: class variable, either 0 or 1

Some columns in the dataset, have the value 0, which is the missing value. The columns with missing values in the given dataset are Glucose, blood pressure, skin thickness, BMI, and insulin. The missing value is treated using the median.

Here, the datatype counts of the variables are:

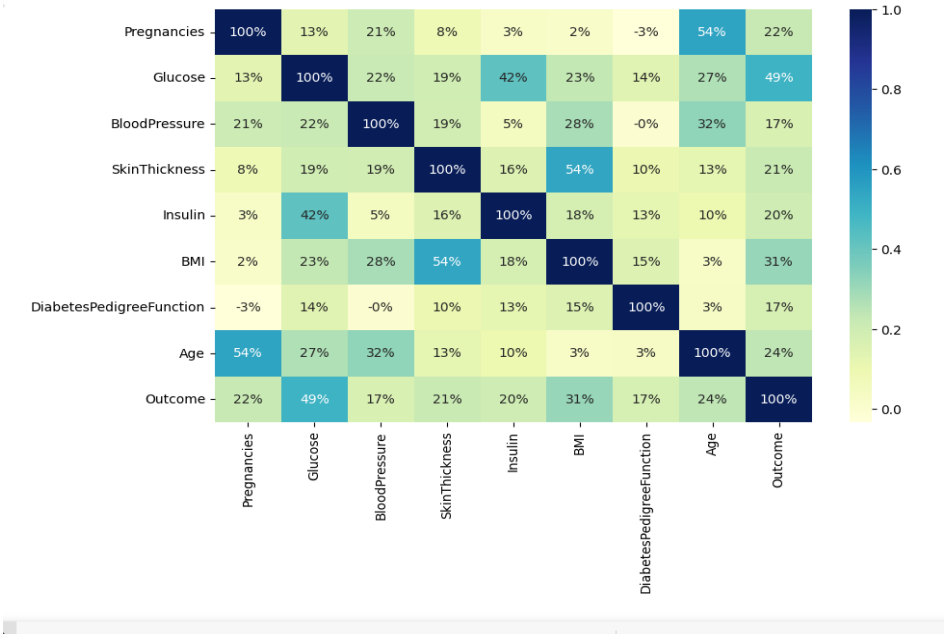


The variables' histograms provide a visual representation of their distributions.



The information gathered from the histograms consists of

- The concentration of glucose is high, between 90-125.
- The blood pressure is high, between 65-80.
- Skin thickness is between 20-40.
- Insulin is high, near 150.
- BMI is high, between 29-37.

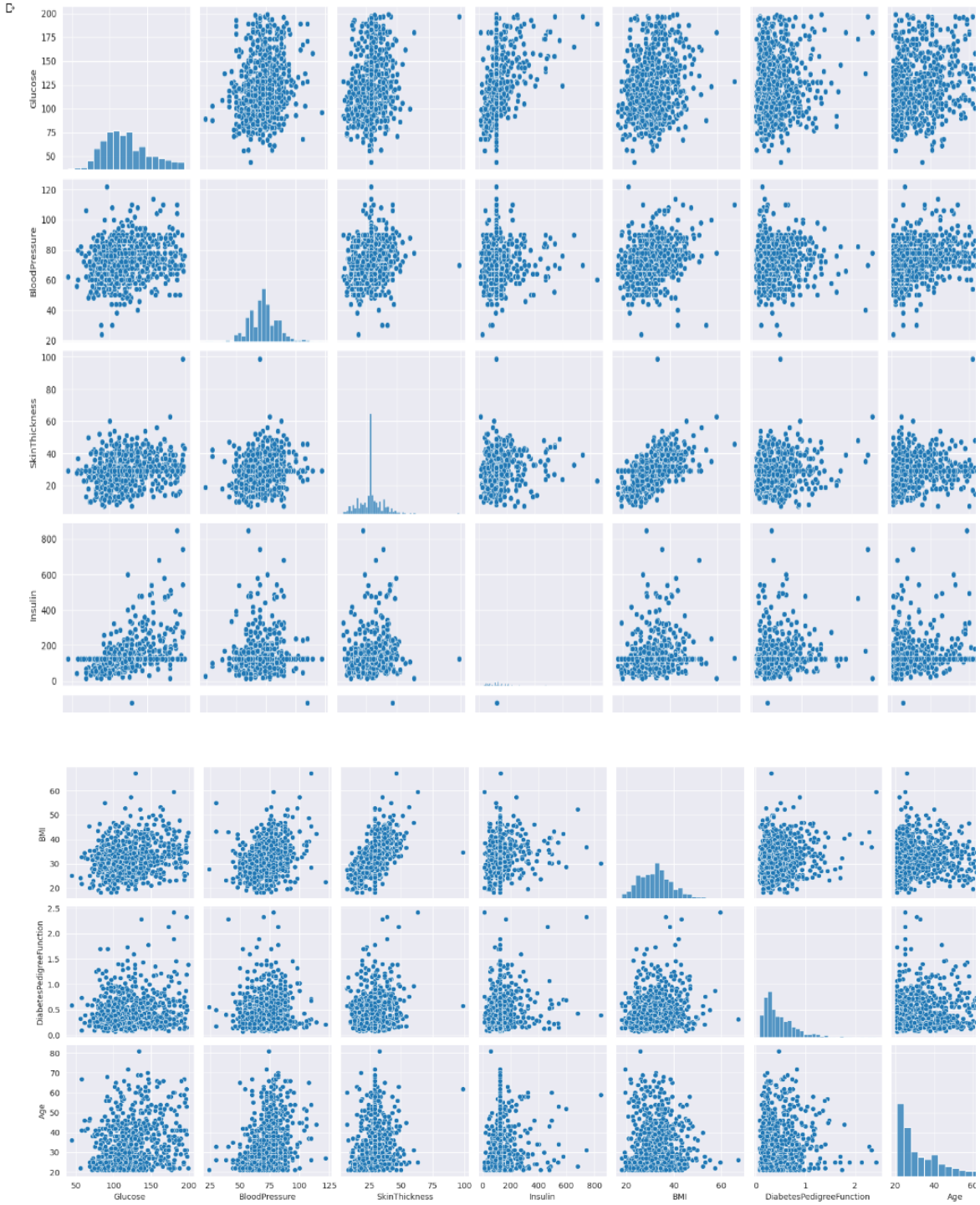


From the heatmap, we can know that age and pregnancies, BMI and skin Thickness, Glucose, and insulin have a good correlation, and diabetic Pedigree Function and pregnancies, blood pressure, and diabetic Pedigree function are negatively correlated.

The outcome column has 268 diabetic (1 patient) and 500 Non-diabetic (0 patient) patients. The column is not balanced, so we do Smote analysis to balance the data.

Scatter-plot:

We can learn about the relationship between the variables using a scatter plot.The Majority of the Scatter plot has a Positive trend. Some outliers can also be observed in the scatter plot. A Strong positive relationship can be seen between the variables BMI and skin thickness, Glucose and insulin.The relationship between insulin and blood pressure, and age and BMI can be observed as scattered.


Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age

We construct a classification model to determine if the patient is diabetic or not. To build the model, we divide the data into training and testing; the training data is utilized to train the algorithm, and the testing data is used to check the performance of the model. However, with cross-validation, we compare different machine learning methods. Here, is the accuracy score of different machine learning classification algorithms.

The accuracy score achieved using Logistic regression was 75.5%

The accuracy score achieved using Decision Tree Classifier was 73%

The accuracy score achieved using SVM was 76.75%

The accuracy score achieved using KNN Classifier is: 76.5 %

The accuracy score achieved using XGBoost is: 82.0 %

Comparing the accuracy score, we can see that XGBoost and SVM Classifier have done better than the others for this particular dataset.

Model performance, i.e., on testing data, can be known from the confusion matrix.

```
Here,

True Positive -> 72

False Positive-> 19

False Negative-> 30

True Negative-> 79

sensitivity = 0.7058823529411765

specificity = 0.8061224489795918
```

Sensitivity is known as the true positive rate or recall.

```
Sensitivity = True Positives/(True Positives + False Negatives)
```

Specificity is known as the true negative rate

```
Specificity = True Negatives/(True Negatives + False positives)
```

By analyzing the confusion matrix, we can derive important performance metrics such as the true positive rate and the false positive rate. These metrics are useful in constructing a ROC (receiver operating Characteristic) curve. The ROC curve provides insights into the model's ability to distinguish between positive and negative classes by varying the threshold value.

By plotting the ROC curve, we can visualize the trade-off between the true positive rate and the false positive rate at different threshold values. It aids in determining an optimal threshold value that balances the classification accuracy for positive and negative instances. The optimal threshold values for the different algorithms are as follows:

For Logistic regression, the optimal threshold is 0.35441014470216464

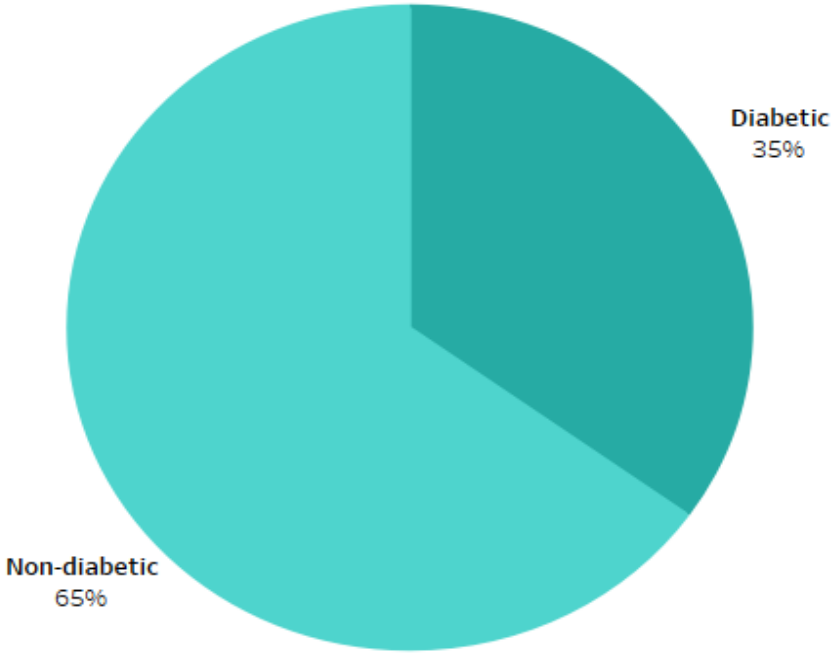
For svm Optimal threshold is 0.32269641782316993

For DecessionTree optimal threshold is 1

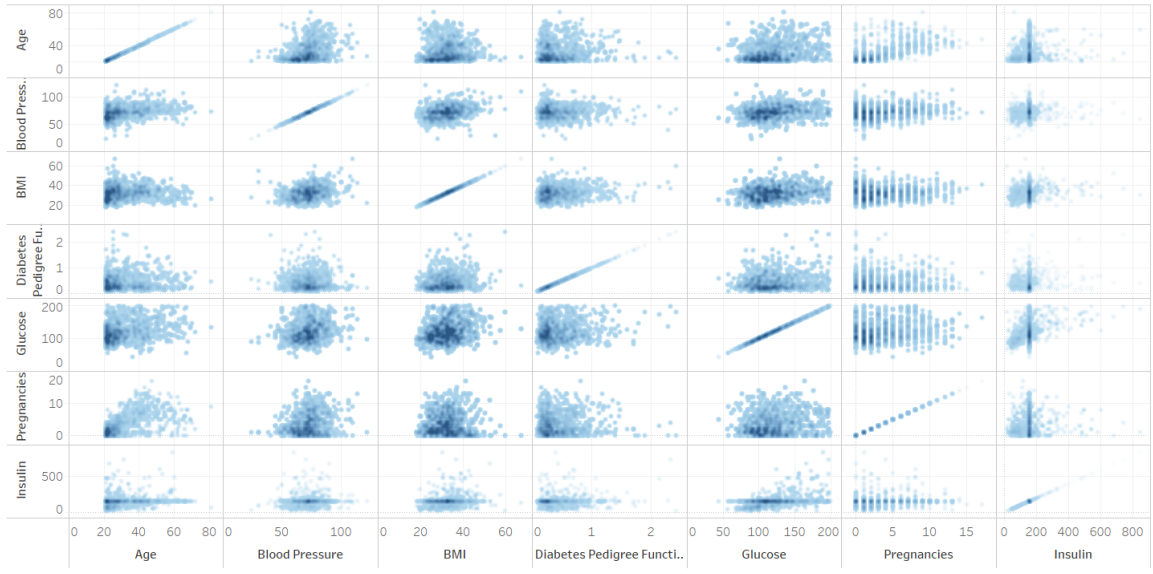
For XGboost Optimal threshold is 0.5294277

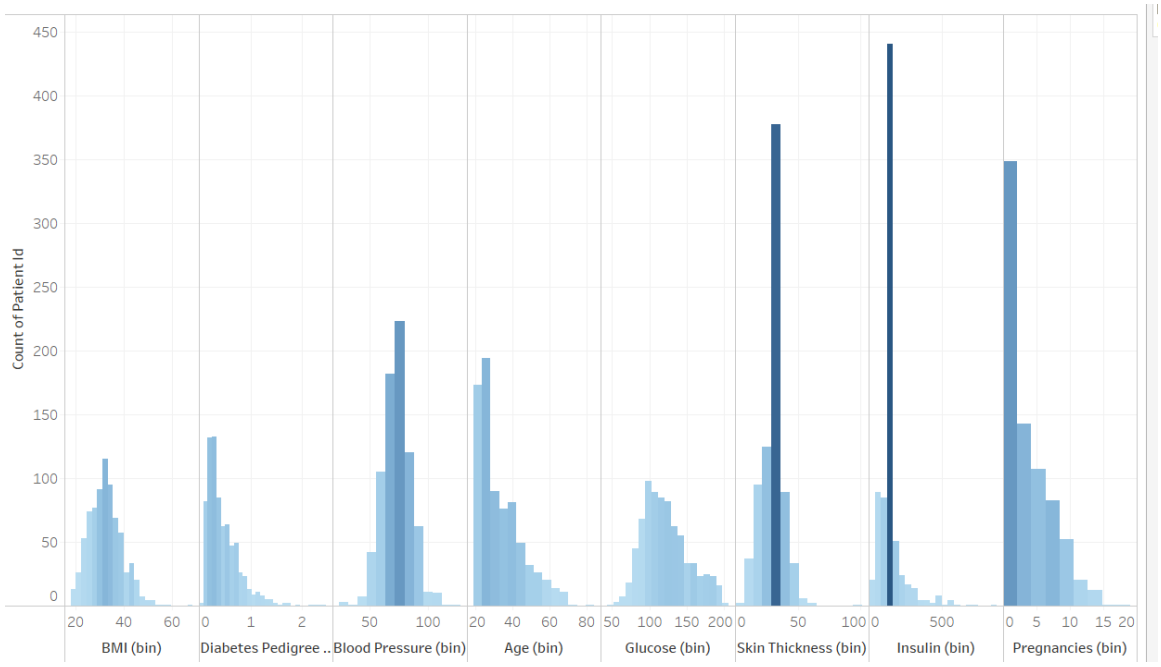
The AUC-ROC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). AUC provides an aggregate measure of performance across all possible classification thresholds. The AUC-ROC curve reveals that XG BOOST outperforms other classifiers at all threshold levels.

Hence, the tableau Dashboard contain the following worksheets:



scatter-plot





heatmap

Column (pivot d (1).cs..	Age	BloodPressure	BMI	DiabetesPedigre..	Column Glucose	Insulin	Outcome	Pregnancies	SkinThickness
Age	1.000	0.325	0.026	0.034	0.267	0.137	0.238	0.544	0.128
BloodPressure	0.325	1.000	0.281	-0.003	0.218	0.073	0.166	0.209	0.193
BMI	0.026	0.281	1.000	0.153	0.231	0.167	0.312	0.022	0.542
DiabetesPedigreeFun..	0.034	-0.003	0.153	1.000	0.137	0.099	0.174	-0.034	0.101
Glucose	0.267	0.218	0.231	0.137	1.000	0.420	0.493	0.128	0.193
Insulin	0.137	0.073	0.167	0.099	0.420	1.000	0.214	0.056	0.158
Outcome	0.238	0.166	0.312	0.174	0.493	0.214	1.000	0.222	0.215
Pregnancies	0.544	0.209	0.022	-0.034	0.128	0.056	0.222	1.000	0.083
SkinThickness	0.128	0.193	0.542	0.101	0.193	0.158	0.215	0.083	1.000

bubblechart