VISUALIZING STATISTICAL ANALYSIS OF MULTI TABULAR ATTRIBUTES WITH SQL

By

VAISHNAVI KAMASANI

Bachelor of Technology

GITAM University

Visakhapatnam, Andhra Pradesh, India

2013

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2016

VISUALIZING STATISTICAL ANALYSIS OF MULTI TABULAR ATTRIBUTES WITH SQL

Thesis Approved:

Dr. Ronak Etemadpour

Thesis Adviser

Dr. Christopher Crick

Dr. David Cline

ACKNOWLEDGEMENTS

We acknowledge Caleydo team for giving the musician dataset used in the domino model for our approach.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: VAISHNAVI KAMASANI

Date of Degree: MAY, 2016

Title of Study: VISUALIZING STATISTICAL ANALYSIS OF MULTI TABULAR ATTRIBUTES WITH SQL

Major Field: COMPUTER SCIENCE

Abstract:

Data extraction and data management is playing a vital role in today's world. Databases are widely used by all the organizations. Analysis of data is very crucial when comparisons are done between different subjects. There are many software's developed for statistical analysis of data. Various visualization techniques are used for representation. In statistical analysis of tabular data in databases, data is either extracted as external sheets or the statistical software's are connected to the servers to test data. In our approach, we introduce a web based interface where users can select any number of attributes and view the results with some simple visualization. SQL queries are written for different methodologies to analyze data. Formulas and structure of all the queries are visualized and represented for the users to understand the query processing and the test methodologies. All the statistical tests are performed on multi tabular data. Ranking is performed on categorical data to replace these values with ranks. With the selected attributes, views are created in the database with the ranks replacing the categorical values in these views. The developed interface is tested with different users to evaluate the visualizations used and the understandability of the statistical tests.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Databases are very essential in today’s organizations. Query processing is used in a wide range for data extraction and modification. Multiple tables are used containing various combinations including views, constraints etc. When large databases are to be analyzed statistically, either software’s are used or the data is extracted externally to analyze the tabular data. To avoid this query processing is used for analysis and queries are written for all the tests and methodologies. All the statistical tests are visualized to make the test structure understandable for the users.

In Section 1, we discussed about the previous work determining the query visualization techniques and tools, which are similar to our approach. This section also shows the advantages and disadvantages of using the tests and other methodologies used in the process. In Section 2, the contribution is explained which shows the improvements and the modifications of the existing and the present research work explaining the ranking process and the tests used for the statistical analysis of the multi tabular subjects. In Section 3, the approach is discussed which explains the structure of the various test queries and shows the relation between the attributes with a network diagram and the query structure is represented as tree. Heat map is constructed which shows the relation between every pair of attributes. In Section 4, the results are shown where a musician dataset is considered which has three tables. The Caleydo team gives this dataset. They developed a “Domino” tool in which this dataset is used to visualize all the attributes.

Ten attributes are considered in this dataset, which are compared using the tests and methodologies mentioned in Section 3. In 4.2, sample queries are given with the nodes connecting graph and the formula representing the queries. In Section 5 we show the user study of twelve people. Here we evaluated the visualization and the understandability of the developed work for any user to easily understand the statistical tests.

In our work we used the web-based interface to make it user friendly for the selection of the attributes. This interface is also made interactive to show the labels and the values in the visualized queries. MySQL queries are written for the extraction of the data as a test result. One test may contain more than one query. Few tables with significant values are inserted from which values are retrieved and these are compared with the calculated query values to get the relation between attributes. Fig 1 shows the overview of the web-based interface with two selected attributes performing T-test and displaying the selected attributes in the network graph with the t value calculation formula.

To perform the appropriate test, the type of data and the distribution of the data should be first checked. The data is either categorical or continuous and the distribution is normal or ordinal. The scatter plot for each attribute is constructed which shows the data is distributed normally or non-normally. If the selection of data is more than two attributes "Anova" test is followed for normal data and "Friedman" test is followed for ordinal data. After these two tests are performed, depending on the type of data, test is performed between every pair of data. If categorical data is compared with categorical or continues data, "Chi Square" test is performed and if numerical data is compared with numerical data "Student T" test is followed. These test values are used to construct a network graph and the connections thickness between the nodes show the p value. Greater the thickness lesser the p value. Smaller the p value, stronger the relation between the attributes. If only two attributes are selected, chi square test or t test are performed depending on the type of data.

The developed interface is a web-based interface that can be accessed in any web environment. Users can select any number of attributes to find the relation between the selected attributes. Heat map shows the correlation between each pair of attributes. Normally distributed data and the non-normally distributed data can be differentiated by the histograms and the normal curves. The developed network graph and the node connection graph shows the structure the statistical test produced. The visualizations developed allow the users to understand the structure of the respective tests. Formulas are produced with on click, mouse over and mouse out events. With all such characteristics our approach can be used in various real world scenarios. The node connection graph, which shows a detailed description and the flow of all the statistical tests, performed. This is very useful for understanding the tests easily. Thus our approach can be used in the educational institutes for teaching the statistical tests, which gives a better understanding to the users. This can also be used in any inline teaching sites, which helps the users to easily understand the flow of the tests. The data analysts can also use this approach, which helps them understand the test procedures and the relations. Since the interface runs for multi tabular attributes, no extraction is needed to work with external data. Advantages include the use of large databases directly with the interface without any software's.

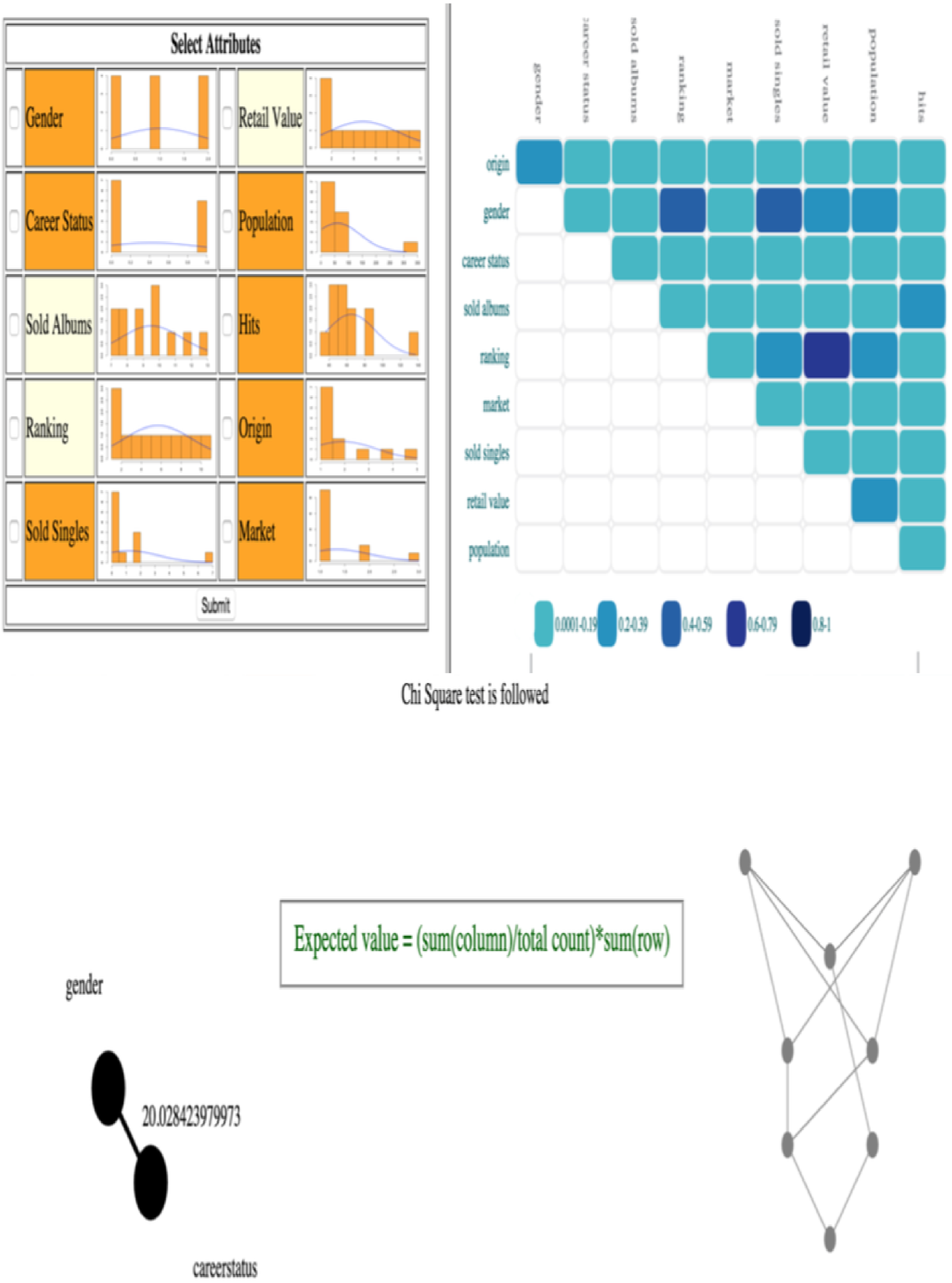


Fig 1: Web Interface representing attribute selection, heat map with correlations between attributes, nodes connection representing the chi square test, network graph between selected attributes and formula of T-test when clicked on the node of tree

CHAPTER II

SECTION 1 (PREVIOUS WORK)

There are various research works performed on databases and statistical analysis daily. Structuring of database queries has been a widely performed research question. Mentioned below are few works similar to our approach and few of the existing works.

Domino [1] tool developed by Caleydo uses multi tabular data for data extraction, comparison and manipulation. This tool also allows the creation of new tables with the existing tables by the user on the dashboard. Data is represented depending on the datatype with partitioned clocks; numerical blocks and matrix block which is a combination of the partitioned and the numerical blocks. 2D heat map is used to represent the hits, the histogram is drawn which represents the number of studio albums and parallel co-ordinates are drawn to show the connection between the attributes. In domino the relation is shown only between two attributes. Our approach is an extension to this, which shows relation between more than two attributes. In domino the relation is only shown to describe the connection between the attributes but do not show any relation between them.

VLDB (Very Large Databases) [3] is using a structural way of representing queries. Every step of the query is represented as a table including the contents of the database tables. This work appears similar to our structuring of the query for visualization. But this visualization is useful when the queries are correlated. If joins or separate queries for same procedure are to be visualized this approach of visualization is not appropriate to use. Complex queries can also make the visualization confusing because in this approach every attribute is displayed with the values in it. In case of large datasets this could not support the visualization used. In this case the nodes connection graph representation can be helpful for better understanding.

Yongjoo Park in Visualization-Aware Sampling for Very Large Databases [10] is using correlation approach for calculation of the amount of data present in the dataset. This was used specially for the geo tagged data. Correlations are found between any two attributes and depending on the value the size is estimated. Greater the value larger the dataset. This is similar to our approach, correlation is used in both the approaches for the data but in this case it was used for the size of data and in our approach the correlations are used for the comparison of the relation between the data. Greater the value stronger the relation in the attributes and smaller the correlation value lesser the relation.

Exploring DATA Step Merges and PROC SQL Joins [11] is published by SAS Global Forum, which shows Venn diagram representation. Though Venn diagrams are very familiar, SAS used their representations to visualize joins in SQL queries. Venn diagrams are used for representation of the common data between to subjects. These are best used for the representation of the joins and sometimes-correlated queries. This representation could be little confusing if the queries are large or complex. If any process or mechanism has to be represented with more than two queries, this representation can be a ciaos to visualize. The nodes connecting graph representing the queries can solve this because it has the connection between nodes with the attributes from multiple tables and the queries continue from these nodes. So these nodes and connections can be used to represent multi queries in the same visualization.

imMens: Real-time Visual Querying of Big Data [16] is a tool developed by the Department of Computer Science, Stanford University which used fragments partition in visualization to represent the data sets. Data cubes are converted to partitioned data tiles for easy modification of the extracted data. Algorithms are built to manipulate the data formulas are return to segregate the data into fragments. This approach for the visualization represents the number of attributes in each attribute and the dimensions of the fragmented cube depend on the number of attributes combined for every fragment. The selection procedure is different with our approach but the selected attributes are used for the visualization in both the techniques.

Interactive visual summarization of multidimensional data [18] published by Systems, Man and Cybernetics used the visualization is presented using a nodes connecting graph again but this gives a complete summary visualization generated with nodes representing attribute value ranges. Larger nodes represent value ranges occurring more often in the extracted association rules. Lines connect value ranges occurring together in the association rules. In the query processing we used, we calculated the values, which are used in the network graph as the thickness of the connection between the nodes.

NakeDB: Database Schema Visualization [17] is a tool, which represents the queries with ER diagrams, radial tree layout, node link tree layout, circular layout, force directed layout and flowcharts. In our approach we used the nodes connecting graph to represent the queries and the query processing. The nodes represent few attributes of the table and the child nodes represent the connection with the other attributes. In the nodes connecting graph we also used the labels and the formulas to represent with query.

Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data [19] was published by Visualization and Computer Graphics, IEEE is similar to the Domino [1] model mentioned previously. The visualization used for both the publications is similar but the data used vary. Domino used the multi tabular data and Interactive Visual Discovering of Movement Patterns with Sparsely Sampled Geo-tagged Social Media Data.

Polaris: A System for Query, Analysis and Visualization of Multi-Dimensional Relational Databases [20] published by proceeding INFOVIS 2000 is similar since it has a user interface but this is not a web interface. It has the dragging and dropping of different elements into the interface which can allow the users to choose their own visualization for any representation. There are also different visualizations representing the dataflow of the processing. This is also given as option to the users. Since this is a tool for visualization and users have an option of creating their own visualization with undo and redo facility.

CHAPTER III

SECTION 2 (CONTRIBUTION)

The points below describe our approach developing the web interface,

- We developed SQL queries for various statistical analysis methodologies and connected with the web interface to make interaction easy for users and simple visualization techniques are used to reduce complexity. This approach can be carried out for any number of subjects irrespective of type of data and distribution of values.
- Categorical data is converted into ranks making the values suitable for query analysis. Any data set with multiple tables can be considered for analysis since the chosen attributes from all the tables are combined to form views. These views contain the updated ranks for the categorical subjects. All the further analysis is performed on this view.
- Network graph is drawn representing the relation between the attributes. Network graph is constructed to show the relation between the selected attributes.
- Nodes connecting graph is given to the query processing and mouse operations display the formulas used at every step of the query. Nodes connecting graph is used to visualize the query processing for every test and mouse operations are used to show the labels and the formulas used in the procedure.

- Differentiation between the distributions of data is represented as scatter plot for every attribute. The analysis carried out can be performed for any number of attributes and on any number of tables.
- The thickness of the connection between the attributes, which are represented as nodes, signifies the p value in the network graph. The thickness is inversely proportional to the p value. Lesser the p value stronger the relation between the attributes.
- Evaluation is conducted on the visualization and the understandability of the statistical tests by giving tasks to 12 users. Significance is calculated with the results of these 12 users with statistical analysis between the values found for accuracy, confidence, time, complexity and efficiency.

CHAPTER IV

SECTION 3 (APPROACH)

As mentioned, this section describes our approach mentioning the tests and methodologies used for developing the interface. Section 3.1, Experimental setup describes the architecture and construction of the web interface. This section also describes the platform used to develop the interface. Section 3.2, overview of the interface describes the operation of the web interface. Section 3.3, data description section describes the dataset used for the approach and the type of data and the data distribution for each attribute. Section 3.4, ranking of data section describes the procedure of converting the categorical data into the numeric and placing them in the views. Section 3.5, selection of the test describes the use of the appropriate test for the appropriate data. Section 3.6, Distribution of data describes the normal and non-normal data scatter plot of each attribute. Chi square test (Section 3.7), T test (Section 3.8), Anova test (Section 3.9), Friedman test (Section 3.10), Pearson's coefficient correlation (Section 3.11) are described in the next sections with the queries.

3.1 EXPERIMENTAL SETUP

Since this is a web-based interface, MVC [13] architecture is followed with MySQL as backend, PHP for the middle ware giving connection between the database and front end, HTML and CSS are used for front-end design with java script and PHP for validations. HTML, SVG and JavaScript are used for the visualizations of network graph and the query nodes connecting graph.

On mouse over, on mouse out and on click actions are performed on the graph to display the formulas used in the queries and other labels. Heat map [8] structure is obtained from Zing Charts and Scatter Plot [9] structure is taken from Am Charts. csx server is used for database tables.

3.2 OVERVIEW OF WEB BASED INTERFACE

Fig 1 shows the user web interface. This is divided into three sections. The first one is for the user to select the desired number of attributes. After submit, in the third section the query structure is displayed which with the network graph which shows the relation between the attributes. The thickness of the connection between two nodes is based on the p value. Less the p value thicker the connection. On mouse over and on mouse out operations can be performed on the query structured graph and the network graph to see all the values, labels and the formula used in the query processing. The second section shows the heat map, which shows the correlation between all the attributes. On mouse over on the blocks of heat map shows the correlation coefficient value. A popup is popped if submit button is clicked without any selection of attributes or if there is no relation between the selected attributes. In the selection region the difference in color of the attributes shows the distribution of the data. The darker color represents the ordinal data whereas the light color represents the normal data. When clicked on the graph, the maximized version of the graph with the scatter plot of the attributes is displayed. Fig 2 shows the network graph between eight selected attributes.

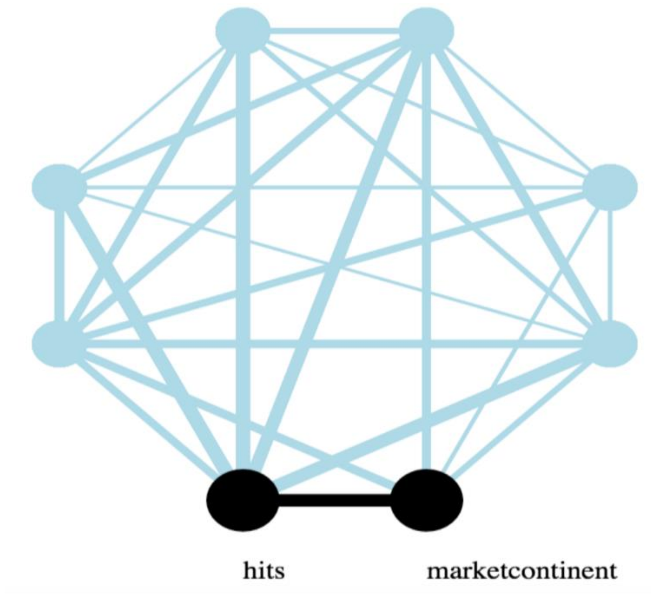


Fig 2: Network graph showing relation between 8 attributes performing Friedman test followed by chi square and t test between each pair of attributers

3.3 DATA DESCRIPTION

In our approach musician data set is used from the Domino [1] model. This dataset has three tables. Ten attributes are considered with twelve artists and the properties and details describing each artist like the attributes. These attributes belong to the multi tabular dataset so views are created with the selected attributes. In these views the categorical values are replaced with the ranks. Table 1 shows the division of the categorical and the numerical values with the data distribution.

Attribute Name	**Data Type**	**Data Distribution**
Gender	Categorical	Non Normal
Career Status	Categorical	Non Normal
Sold Albums	Numerical	Normal
Ranking	Numerical	Normal

Sold Singles	Numerical	Non Normal
Retail Value	Numerical	Normal
Population	Numerical	Non Normal
Hits	Numerical	Non Normal
Origin	Categorical	Non Normal
Market	Categorical	Non Normal

Table 1: Attributes data type and distribution

3.4 RANKING OF CATEGORICAL DATA

Analysis of data is performed between numerical and numerical, numerical and categorical, categorical and categorical groups. Since all the analysis is formula based categorical data is converted into ranks depending on the count of data. Count of distinct variables are calculated and ranked based on the maximum count. The values are then replaced with these ranks in the views created from these tables containing the selected attributes. If the count values are same for different distinct values, ranks are given randomly to these variables and analysis is performed. These ranks are now considered as different groups and not as numerical values. If the count value is same for more than two subjects, after the randomization of ranks, analysis is performed for all the combinations exchanging the ranks for every combination and the possible least p value is considered. Fig 3 explains about the ranking process.

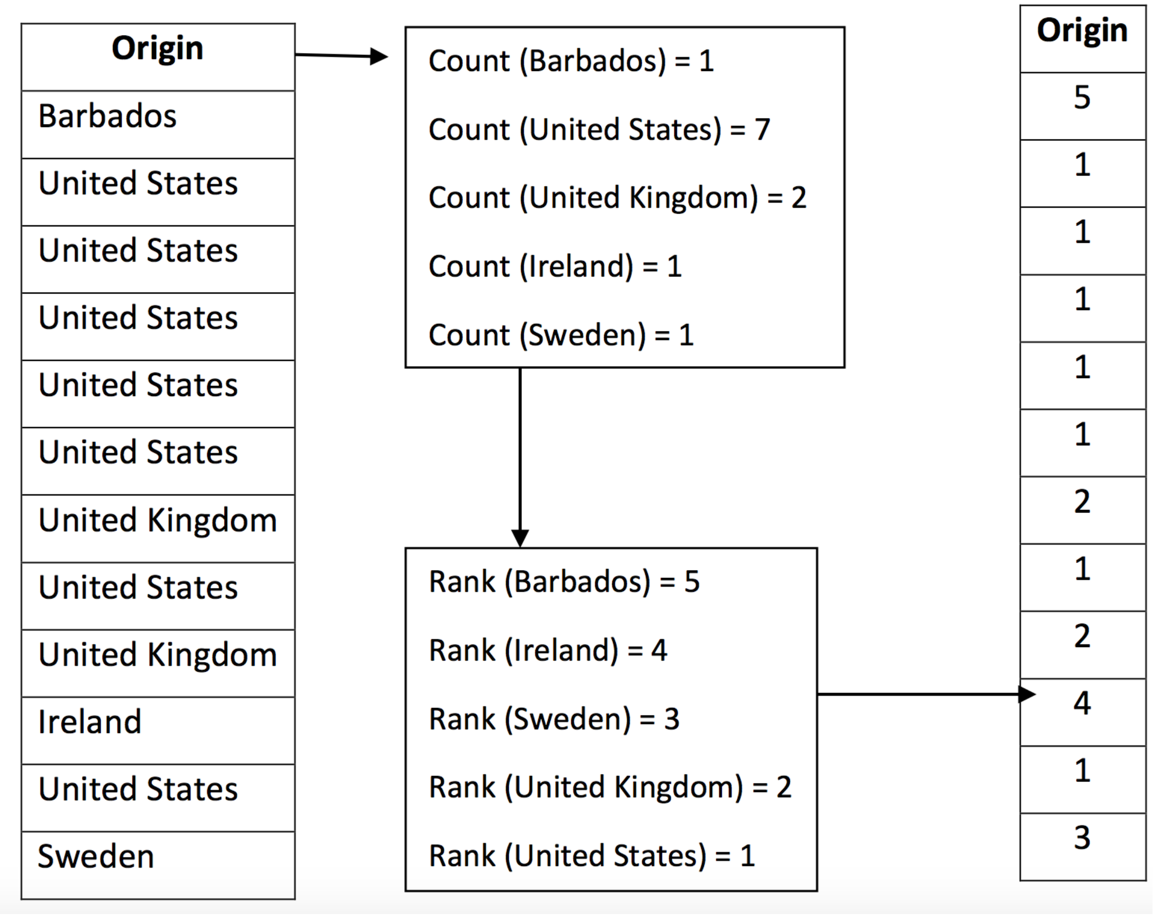


Figure 3: Ranking process to convert categorical data into numerical data

3.5 SELECTION OF TEST

When there are two groups for comparison, "T-test" is performed for numerical data and "Chi-square test" is followed for categorical data. If the selection is made with more than three groups, the data is checked for normality. Nominal data follows "Anova test" and ordinal data follows "Friedman test" irrespective of categorical or numerical values. If the significant p-value is less than 0.05 further calculations take place for analysis between each pair of attributes using T-test or Chi-Square test and a network graph is constructed to represent these relations. "Pearson's Correlation Coefficients" are calculated and represented as a heat map showing the correlation between each pair of attributes. Table 2 shows the selection of the appropriate test for the comparison between types of data. This table shows the comparison for two attributes. Table 3 shows the suitable test for the distribution of data for selection of more than two attributes.

Data Type 1	**Data Type 2**	**Test Performed**
Numerical	Numerical	T test
Numerical	Categorical	Chi Square
Categorical	Categorical	Chi Square

Table 2: Selection of test for appropriate data type

Distribution of Data	**Test Performed**
Normal Data	Anova Test
Non Normal Data	Friedman Test

Table 3: Selection of test for data distribution

3.6 DISTRIBUTION OF DATA

A multi tabular data set of musicians is taken for the analysis of results. This data set contains three tables with different attributes related to twelve musicians. Ten attributes are considered for comparison and tests are performed. These attributes include numerical and categorical subjects with nominal and ordinal data. Depending on the type of distribution respective tests are performed. Fig 4 is the representation of ordinal data. Fig 5 is the representation of nominal data. In the interface when clicked on the graph a new tab opens with the selected histogram data. Shapiro - Wilk test is performed to check the normality of data. “R” software is used to check the normality and the histograms with the normal curves are also constructed using “R”. The normality depends on the p value produced by the Shapiro - Wilk test performed in “R”. If the p value is greater than 0.05, the data is said to be distributed normally. If the obtained p value is less than 0.05, the data is not distributed normally. The normal curve also differ its path.

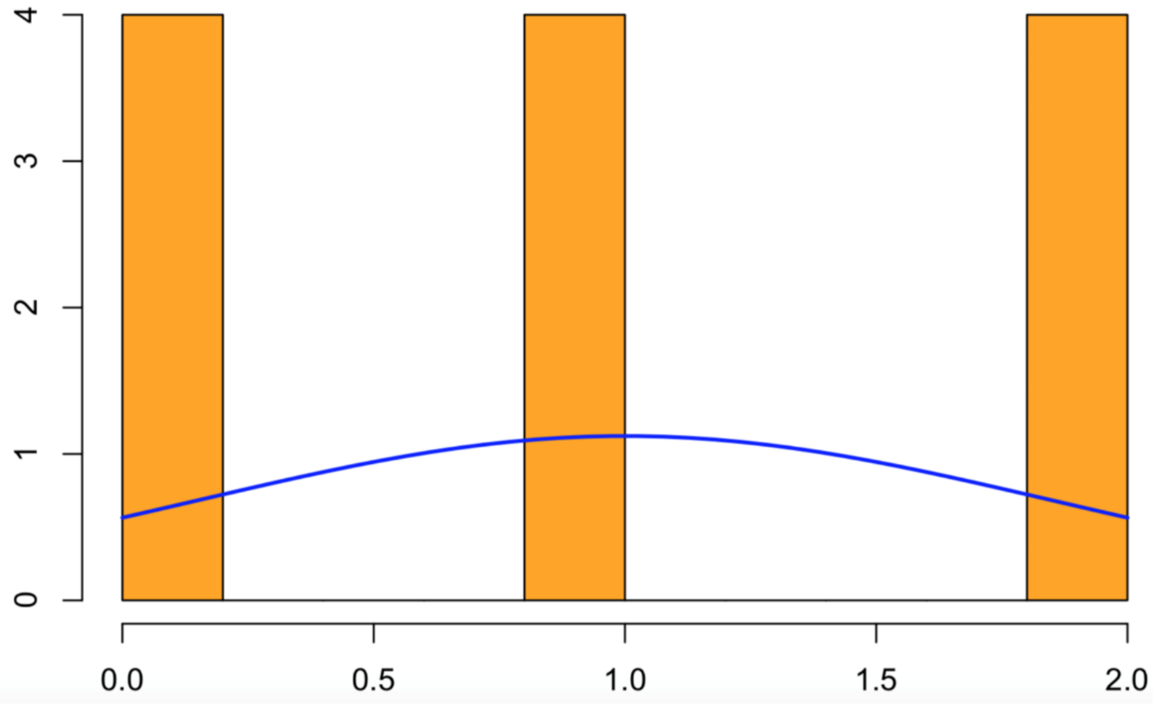


Figure 4: Non Normal data distribution

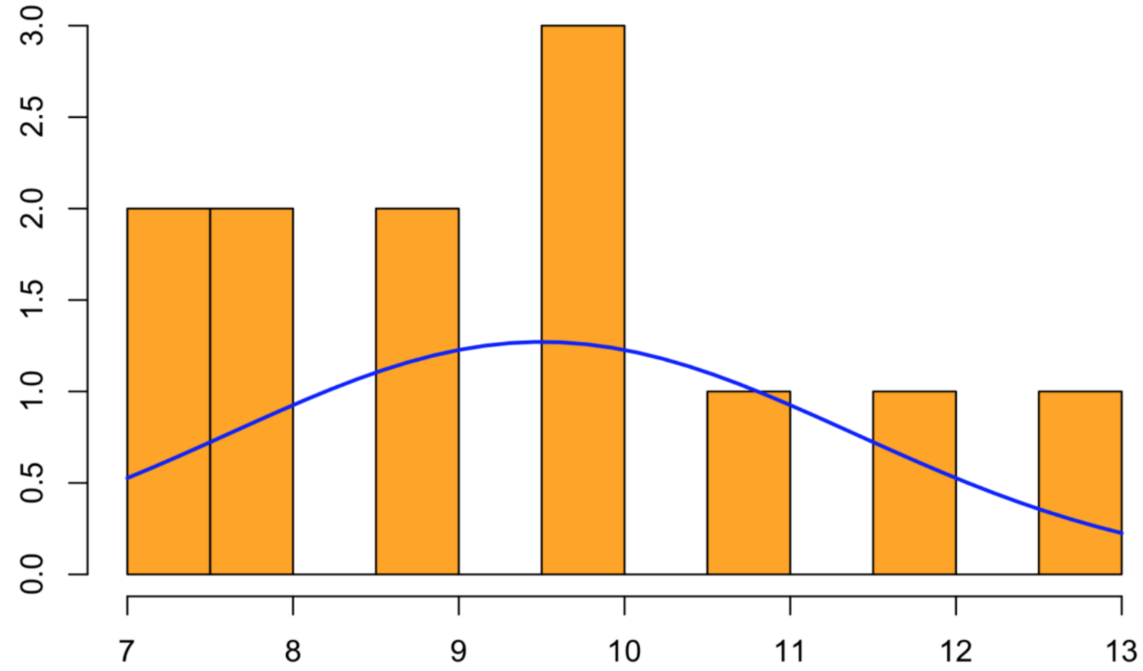


Figure 5: Normal data distribution

3.7 CHI - SQUARE TEST

Chi Square test is performed for the combinations like the categorical and categorical, categorical and numerical groups. The categorical data is converted into ranks and views are created with the ranked values and the selected attributes.

Chi Square test is very advantageous when compared to other tests because this can perform easy computations irrespective of the size of data. Few tests are limited to the size of the dataset [14], so chi square is the best considered when multi tabular data is used.

Formula to calculate chi square value [7]:

$$X^2 = \sum[(\frac{Or,c - Er,c)^2}{Er,c}]$$

Where,

$O_{r,c}$ = Observed frequency at r level in attribute[0] and c level in attribute[1].

$E_{r,c}$ = Expected frequency at r level in attribute[0] and c level in attribute[1].

Σ = Sum of above across all cells

After the calculations, X^2 ratio from the X^2-table at the obtained df ([r - 1]*[c - 1]) value is compared with the X^2 value. If this value exceeds the ratio, p-value is less than 0.05 significance which indicates there is a relation between the two selected attributes else the test is rejected. Fig 6 shows the network graph of the chi square test between two attributes. The thickness of the connection between the nodes describes the relation between each pair of attributes represented as nodes. Larger the thickness, less the p value.

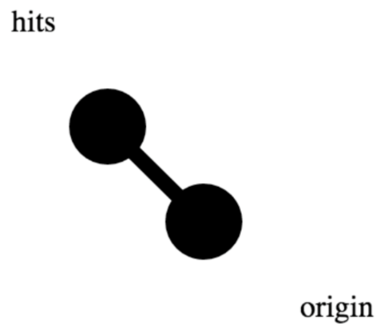


Figure 6: Ci Square network graph

Sample queries for chi square test:

- select (sum((observed–expected)*(observed-expected)))/expected from attributes;
- select @x:=(sum(". $newvarible [0].") / (count (". $newvarible [0].") +count (". $newvarible [1].")))*sum(@y);

3.8 T TEST

T-test is performed on numerical and numerical combinations of attributes.

There are various tests for comparisons between numerical data. Every test has its own advantages. T test combines the crucial points of all the tests like robustness, ease of calculation, else of gathering data and most importantly the simplicity of interpretations [15], this makes it the best test for comparing two samples of numerical data.

Formula to calculate t value [5]:

$$t = \frac{\mathrm{Mx} - \mathrm{My}}{\sqrt{\frac{\left(\sum X^2 - \frac{(\sum X)^2}{Nx}\right) + \left(\sum Y^2 - \frac{(\sum Y)^2}{Ny}\right)}{\mathrm{Nx} + \mathrm{Ny} - 2}\left(\frac{1}{\mathrm{Nx}} + \frac{1}{\mathrm{Ny}}\right)}}$$

Where,

Σ = sum the following scores

M_x = mean for Group A

M_y = mean for Group B

X = score in Group 1

Y = score in Group 2

N_x = number of scores in Group 1

N_y = number of scores in Group 2

After the calculation, t ratio is obtained from the t-table at the obtained df ($[N_x - 1]*[N_y - 1]$). This ratio is compared with the t value calculated. If this value exceeds the ratio, p-value is less than 0.05 significance which indicates there is a relation between the two selected attributes else the test is rejected. Fig 7 shows the network diagram of T test. The thickness of the connection between the nodes describes the relation between each pair of attributes represented as nodes. Larger the thickness, less the p value.

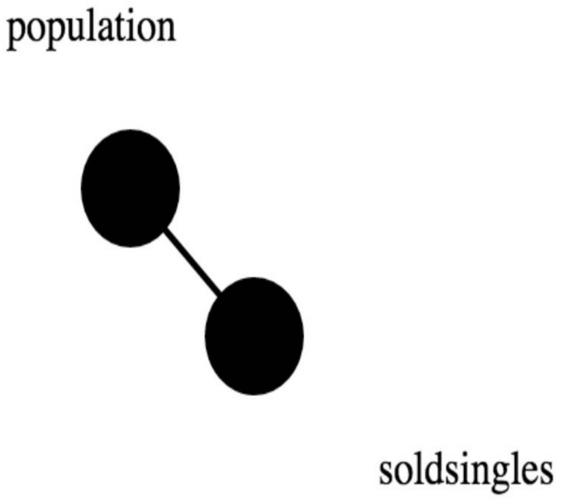


Figure 7: T test network graph

Sample query for t test:

- select abs(((sum(".$newvarible[0].")/count (". $newvarible [0]."))- (sum (". $newvarible [1].")/count (". $newvarible [1].")))/sqrt (((((sum (". $newvarible [0]."*". $newvarible [0]."))- ((sum (". $newvarible [0].")*sum (". $newvarible [0]."))/ (count (". $newvarible [0]."))))+ ((sum (". $newvarible [1]."*". $newvarible [1]."))- ((sum (". $newvarible [1].")*sum (". $newvarible [1]."))/ (count (". $newvarible [1].")))))/ (count (". $newvarible [0].")+count (". $newvarible [1].")-2))*((1/count (". $newvarible [0]."))+ (1/count (". $newvarible [1]."))))) as tratio from attributes;

3.9 ANOVA TEST

Anova test is performed when more than two attributes are selected for analysis. This test is used for the statistical analysis of normal data. If a significant value is found there exists a relation between the attributes, now chi square and t-test are performed on individual pairs depending on the type of data. Anova table is created as a view with different values, which form the anova table. Table 3 represents the anova table with the formulas for the calculation of the f value.

Anova Table [4]:

$$SS = \sum X^2 - \sum X^{2} /N$$

Where,

$\sum$ = sum of the values

X = selected group

N = count of values in the selected group

SS = total value (this is used in the anova table for further calculation)

After the calculations, f ratio from the f-table at the obtained with dfb and dfw values which is compared with the f value. If this value exceeds the ratio, p-value is less than 0.05 significance which indicates there is a relation between the selected attributes else text is rejected. A network graph is produced with obtained t or X^2 values after the anova test.

Source	**Df**	**SS**	**MS**	**F**
Between	no. of subjects-1	$\{\sum_{x=1}^{k}(\sum x)_x^2/nx\} - (\sum X)_t^2/N$	SSb/dfs	MSb/MSw
Within	total count of data - no. of subjects	$\sum_{t=1}^{k} SS^t$	SSw/dfw	

Table 4: Anova Table

The thickness of the connection between the nodes indicates the range of the p-value. Thicker the connection lesser the p value. Fig 8 shows the network graph with three selected attributes showing the relation between them. The thickness of the connection between the nodes describes the relation between each pair of attributes represented as nodes. Larger the thickness, less the p value.

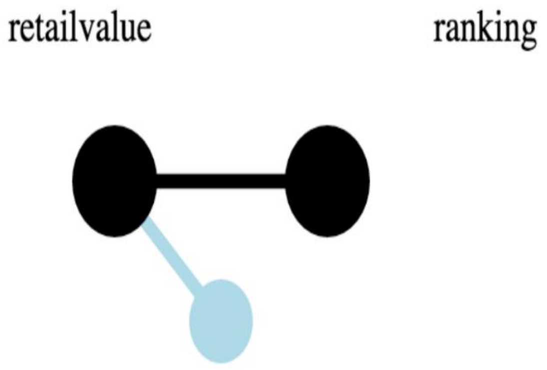


Figure 8: Network graph for anova with 3 attributes

Sample query for Anova test:

- create view details as select ((((sum(".$newvarible[0]."))*(sum (". $newvarible [0].")))/ (count (". $newvarible [0].")))+ (((sum (". $newvarible [1]."))*(sum (". $newvarible [1].")))/ (count (". $newvarible [1].")))+ (((sum (". $newvarible [2]."))*(sum (". $newvarible [2].")))/ (count (". $newvarible [2]."))))- ((((sum (". $newvarible [0]."))+ (sum (". $newvarible [1]."))+ (sum (". $newvarible [2].")))*((sum (". $newvarible [0]."))+ (sum (". $newvarible [1]."))+ (sum (". $newvarible [2]."))))/ ((count (". $newvarible [0]."))+ (count (". $newvarible [1].")+ (count (". $newvarible [2]."))))) as ssb, ((sum(".$newvarible[0]."*". $newvarible [0]."))- (((sum (". $newvarible [0]."))*(sum (". $newvarible [0].")))/ (count (". $newvarible [0]."))))+ ((sum (". $newvarible [1]."*". $newvarible [1]."))- (((sum (". $newvarible [1]."))*(sum (". $newvarible [1].")))/ (count (". $newvarible [1]."))))+ ((sum (". $newvarible [2]."*". $newvarible [2]."))- (((sum (". $newvarible [2]."))*(sum (". $newvarible [2].")))/ (count (". $newvarible [2].")))) as ssw, (((count(".$newvarible[0]."))+ (count (". $newvarible [1]."))+ (count (". $newvarible [2].")))-3) as dfw from tablename;

3.10 FRIEDMAN TEST

Friedman test is also performed if the selected groups are more than two. This test is performed on ordinal data. If a significant value is found there exists a relation between attributes, now chi square and t-test are performed on individual pairs depending on the type of data.

Friedman test calculation:

In Friedman test ranks are taken from each column instead of the data. This is the reason the categorical data is converted into ranks. When random ranks are taken for the same count of distinct values, test is repeated exchanging the ranks between the common count of attributes and least p value is considered for comparison.

T_a = the sum of the n ranks in column A

M_a = the mean of the n ranks in column A

T_b = the sum of the n ranks in column B

M_b = the mean of the n ranks in column B

T_c = the sum of the n ranks in column C

M_c = the mean of the n ranks in column C

T_{all} = the sum of the nk ranks in all columns combined. [In all cases equal to nk (k+1)/2]

M_{all} = the mean of the nk ranks in all columns combined. [In all cases equal to (k+1)/2]

Formula for test [6]:

Substituting the above-calculated values.

$$SS_{bg}(R) = \Sigma\ [n_g\ (M_g — M_{all})\ ^2]$$

Where,

“g” means any particular group and “n” means number of subjects.

Now Chi Square is again calculated using the formula:

$$X^2 = SS_{bg}(R)/\ (k\ (k+1)/12)$$

After the calculation, X^2 ratio from the X^2-table at the obtained df ([r - 1]*[c - 1]) value is compared with the X^2 value. If this value exceeds the ratio, p-value is less than 0.05 significance which indicates there is a relation between the selected attributes else the test is rejected. A network graph is produced with obtained t or X^2 values after the Friedman test. Thicker the connection lesser the p value. Fig 9 shows the network graph with selected seven attributes with the relation between them. The thickness of the connection between the nodes describes the relation between each pair of attributes represented as nodes. Larger the thickness, less the p value.

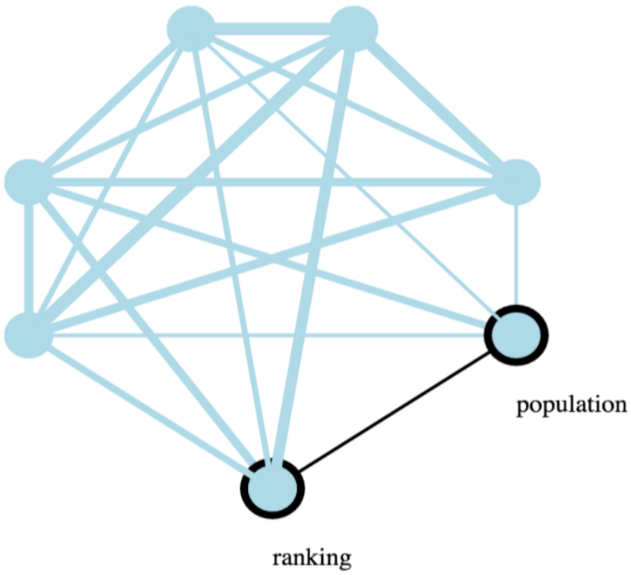


Figure 9: Network graph for Friedman test between 7 attributes

Sample query of Friedman test:

- select (count(".$newvarible[0].")*((sum (". $newvarible [0].")/count (". $newvarible [0]."))- ((sum (". $newvarible [0].")/count (". $newvarible [0].")+ (sum (". $newvarible [1].")/count (". $newvarible [1]."))+ (sum (". $newvarible [2].")/count (". $newvarible [2].")))2) + (count (". $newvarible [1].")*((sum (". $newvarible [1].")/count (". $newvarible [1]."))- ((sum (". $newvarible [0].")/count (". $newvarible [0].")+ (sum (". $newvarible [1].")/count (". $newvarible [1]."))+ (sum (". $newvarible [2].")/count (". $newvarible [2].")))2) + (count (". $newvarible [1].")*((sum (". $newvarible [1].")/count (". $newvarible [1]."))- ((sum (". $newvarible [0].")/count (". $newvarible [0].")+ (sum (". $newvarible [1].")/count (". $newvarible [1]."))+ (sum (". $newvarible [2].")/count (". $newvarible [2].")))2);

3.11 PEARSON’S CORRELATION

Pearson correlation is widely used to find the basic relation between any two groups. These correlations are computed only for numerical data. Since the categorical data are converted into ranks the correlation for categorical groups is found using this data.

Formula of Pearson’s Correlation Coefficient Calculation [12]:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

Where,

N = number of pairs

X Y= product of XY

Σ X Y = multiply each X times each Y, then sum the products

R-value is calculated which ranges from -1 to 1. In our query calculation negative symbol is ignored and values are taken from 0 to 1. The relations between the pairs are represented using a heat map. Fig 10 shows the heat map with the person’s coefficient correlation represented.

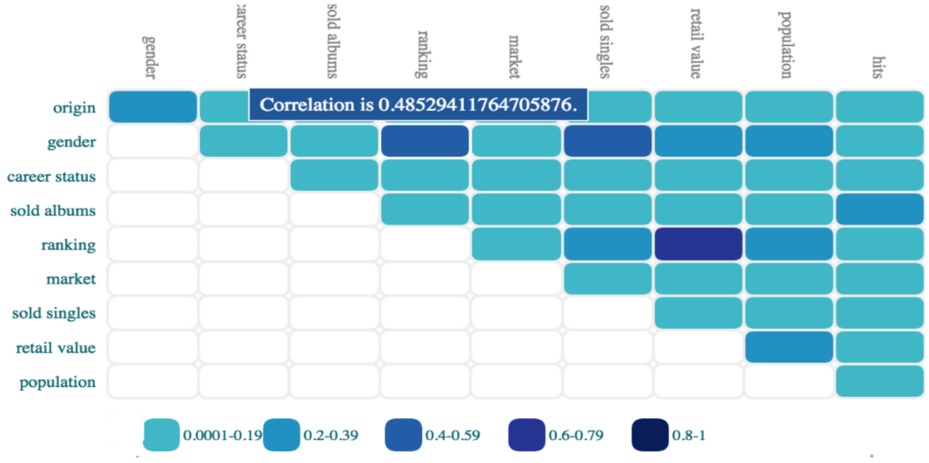


Figure 10: Heat map representing the correlations

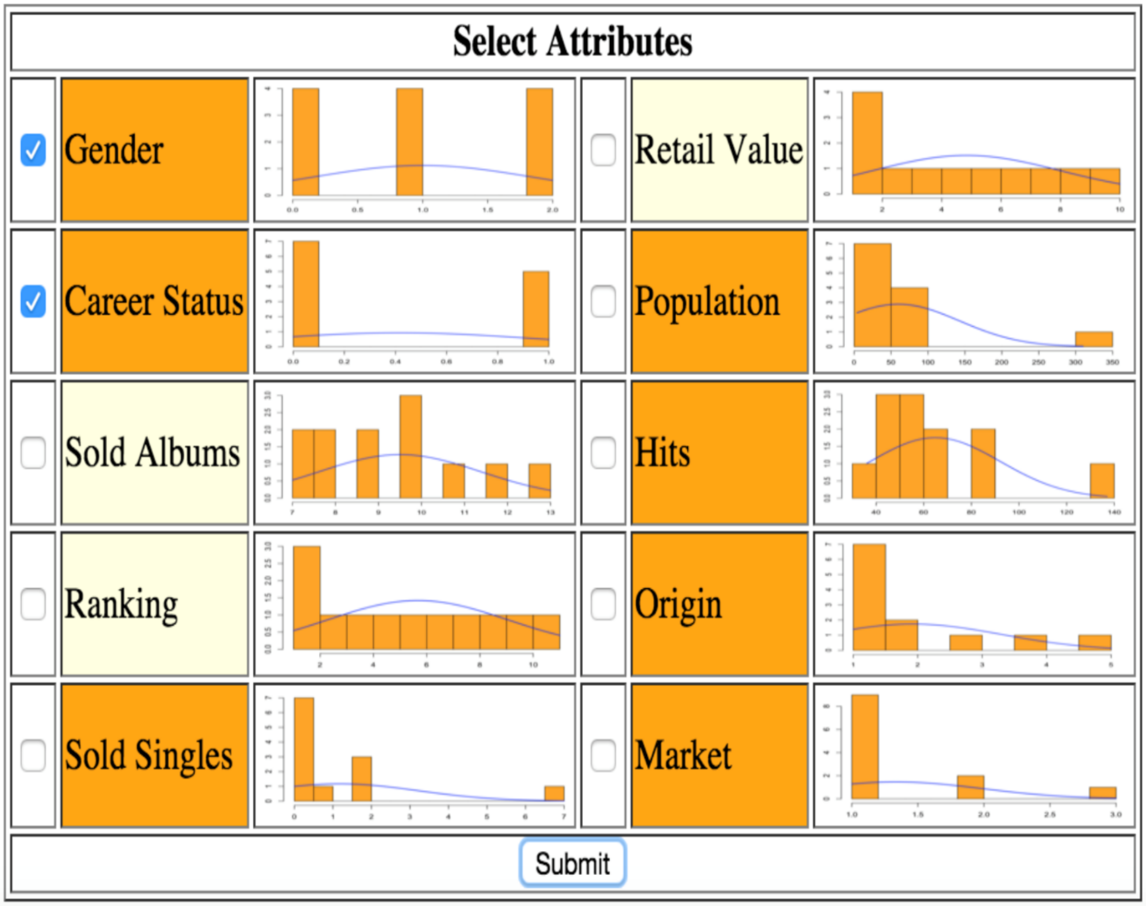
Sample Queries for Person’s correlation:

- Select @x: =avg (". $newvarible [0]."), @y: = avg (". $newvarible [1]."), @div: = (stddev_samp (". $newvarible [0].") * stddev_samp (". $newvarible [1].")) from tablename;
- select ((sum((".$newvarible[0]."-@x)*(". $newvarible [1]."-@y))/ ((count (". $newvarible [0].")-1) *@div))*(sum ((". $newvarible [0]."-@x)*(". $newvarible [1]."-@y))/ ((count (". $newvarible [1].")-1)*@div))) as p from tablename;

CHAPTER V

SECTION 4 (RESULTS)

All the results are performed on the web based interface build with the multi tabular musician data. Fig 11 shows the attribute selection for the tests.



Select Attributes

☑ Gender

☑ Career Status

☐ Sold Albums

☐ Ranking

☐ Sold Singles

☐ Retail Value

☐ Population

☐ Hits

☐ Origin

☐ Market

Figure 11: Attribute Selection

4.1 STATISTICAL RESULTS

If two attributes are selected and the combination is with the categorical variables chi square test is performed on the attributes. Fig 12 shows the nodes connecting graph of chi square test query with the node labels.

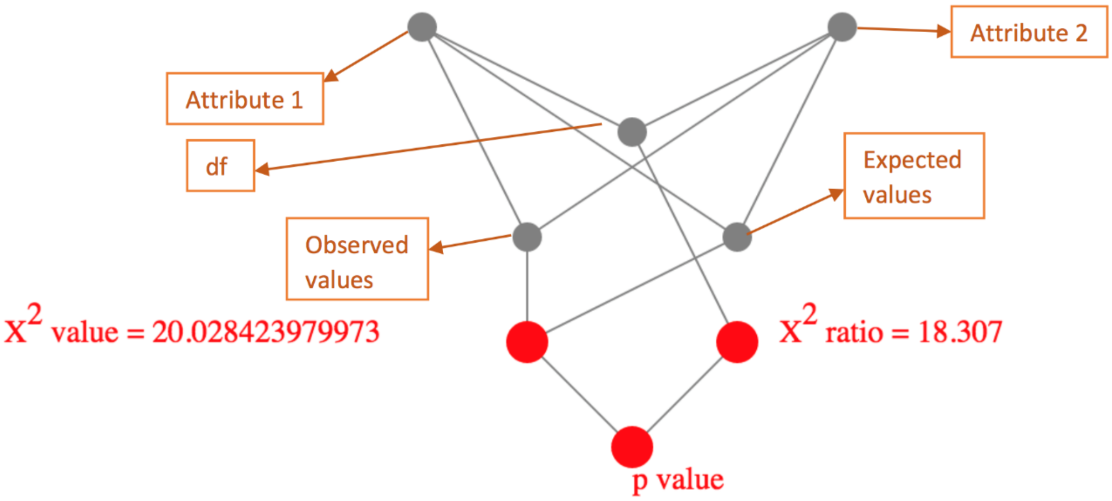


Figure 12: Chi Square test node connection graph (the values in red represent the on mouse over operation displayed in the visualization)

If the two selected attributes are numerical, T test is performed on the attributes. Fig 13 shows the nodes connecting graph of the T test query with the node labels and operations.

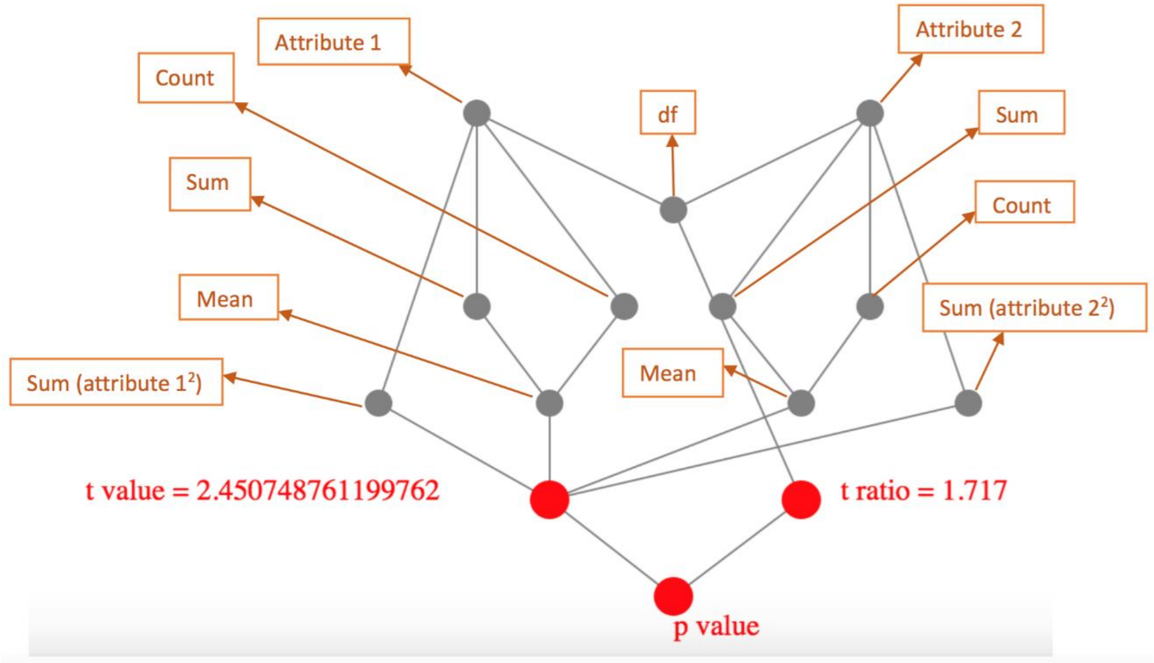


Figure 13: Representation of T test node connection graph (the values in red represent the on mouse over operation displayed in the visualization)

If there are more than three attributes selected depending on the distribution of the data, further test is followed. If the selected attributes are normally distributed Anova test is performed.

Fig 14 shows the nodes connecting graph representation of the Anova test query processing between three normally distributed attributes.

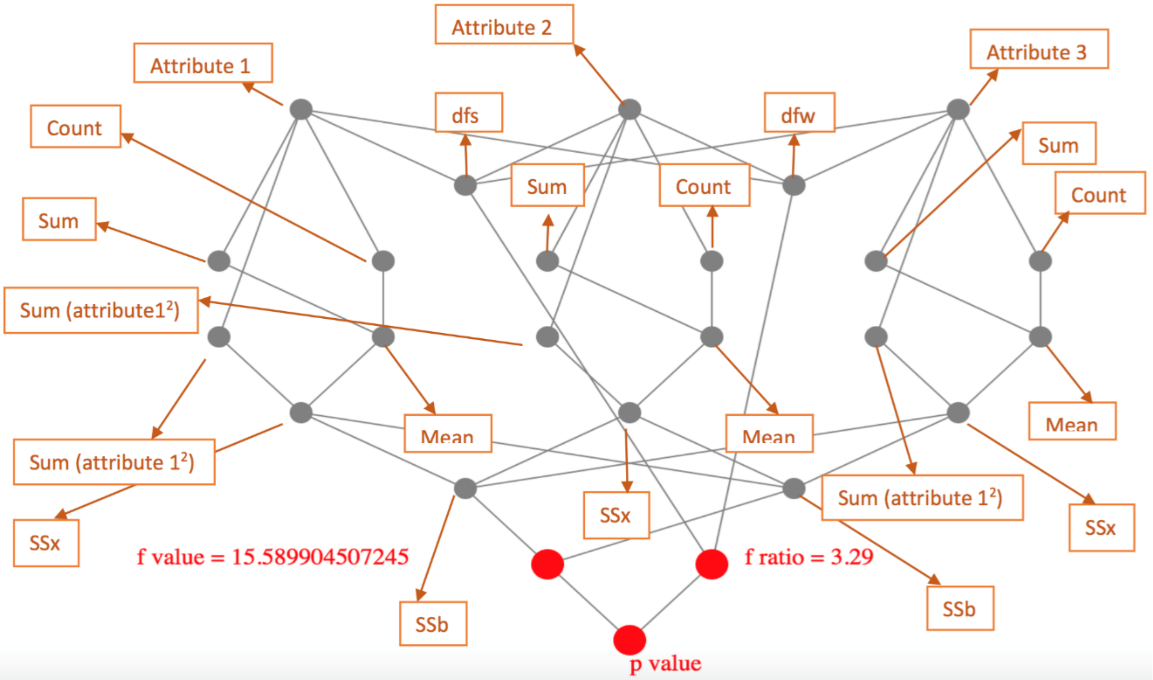


Figure 1: Representation of Anova Test node connection graph (the values in red represent the on mouse over operation displayed in the visualization)

If the selected attributes are not normally distributes Friedman test is followed.

Fig 15 shows the nodes connecting graph showing the query processing of Friedman test for three attributes with the node labels and values.

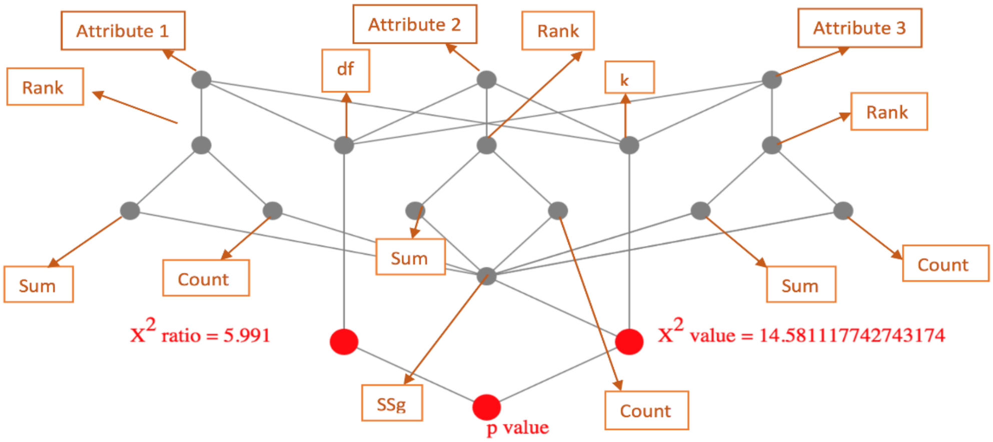
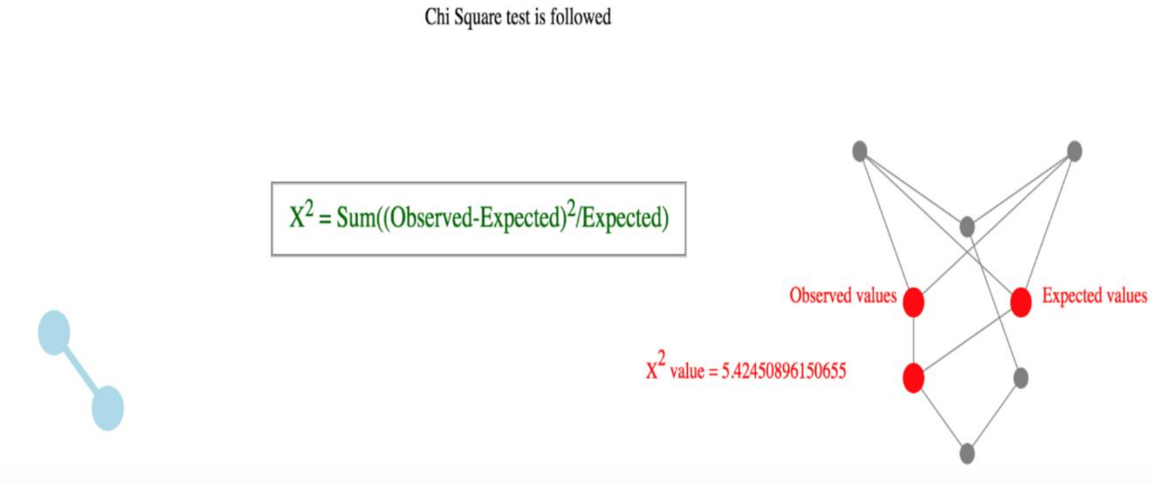


Figure 2: Representation of Friedman test node connection graph (the values in red represent the on mouse over operation displayed in the visualization)

After the query nodes connecting graph is displayed on click action on the mouse shows the formula of the node. Fig 16 shows the chi square value calculation formula on click on the node.

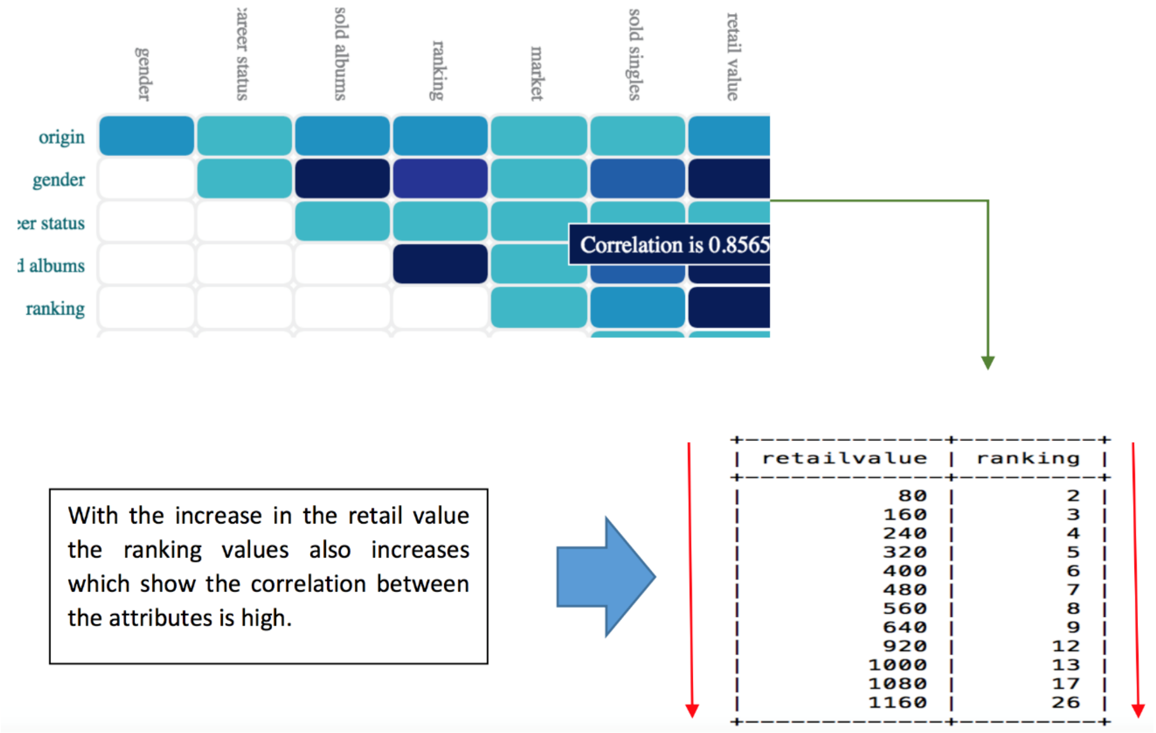


Chi Square test is followed

X^2 = Sum((Observed-Expected)^2/Expected)

Figure 3: Formula representing the X^2 value

Apart from the tests and methodologies performed on the selected data and distribution of values, Pearson’s correlation coefficients are performed on every pair of attributes. Fig 17 shows the correlation coefficients and the values of the two attributes increasing with the strong correlation value.



With the increase in the retail value the ranking values also increases which show the correlation between the attributes is high.

```
+-------------+-----------+
| retailvalue | ranking  |
+-------------+-----------+
|          80 |         2 |
|         160 |         3 |
|         240 |         4 |
|         320 |         5 |
|         400 |         6 |
|         480 |         7 |
|         560 |         8 |
|         640 |         9 |
|         920 |        12 |
|        1000 |        13 |
|        1080 |        17 |
|        1160 |        26 |
+-------------+-----------+
```

Figure 4: Correlation Coefficient between two attributes

4.2 SAMPLE QUERIES

Below are few sample queries used for the test calculations. Fig 18 shows the part of the Anova test query structure with the selected part of the sample Anova query. Fig 19 shows the part of the T test query structure with the selected part of the sample T test query.

- Anova test:

  create view details as select ((((sum(".$newvarible[0]."))*(sum (". $newvarible [0].")))/ (count (". $newvarible [0].")))+ (((sum (". $newvarible [1]."))*(sum (". $newvarible [1].")))/ (count (". $newvarible [1].")))+ (((sum (". $newvarible [2]."))*(sum (". $newvarible [2].")))/ (count (". $newvarible [2]."))))- ((((sum (". $newvarible [0]."))+ (sum (". $newvarible [1]."))+ (sum (". $newvarible [2].")))*((sum (". $newvarible [0]."))+ (sum (". $newvarible [1]."))+ (sum (". $newvarible [2]."))))/ ((count (". $newvarible [0]."))+ (count (". $newvarible [1].")+ (count (". $newvarible [2].")))))) as ssb, ((sum(".$newvarible[0]."*". $newvarible [0]."))- (((sum (". $newvarible [0]."))*(sum (". $newvarible [0].")))/ (count (". $newvarible [0].")))))+ ((sum (". $newvarible [1]."*". $newvarible [1]."))- (((sum (". $newvarible [1]."))*(sum (". $newvarible [1].")))/ (count (". $newvarible [1].")))))+ ((sum (". $newvarible [2]."*". $newvarible [2]."))- (((sum (". $newvarible [2]."))*(sum (". $newvarible [2].")))/ (count (". $newvarible [2].")))) as ssw, (((count(".$newvarible[0]."))+ (count (". $newvarible [1]."))+ (count (". $newvarible [2]."))))-3) as dfw from tablename;

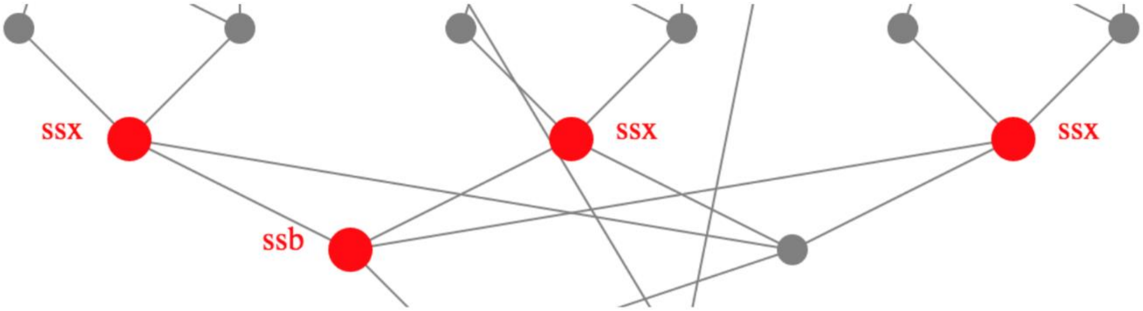


Figure 18: Query and graph showing the calculation of SSb in Anova Test

- T test:

select abs(((sum(".$newvarible[0].")/count (". $newvarible [0]."))- (sum (". $newvarible [1].")/count (". $newvarible [1].")))/sqrt (((((sum (". $newvarible [0]."*". $newvarible [0]."))- ((sum (". $newvarible [0].")*sum (". $newvarible [0]."))/ (count (". $newvarible [0].")))))+ ((sum (". $newvarible [1]."*". $newvarible [1]."))- ((sum (". $newvarible [1].")*sum (". $newvarible [1]."))/ (count (". $newvarible [1].")))))/ (count (". $newvarible [0].")+count (". $newvarible [1].")-2))*((1/count (". $newvarible [0]."))+ (1/count (". $newvarible [1]."))))) as tratio from attributes;

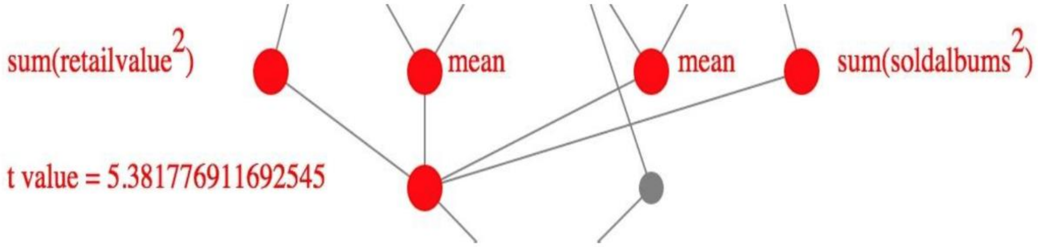


Figure 5: Query and graph showing the calculation of t value in T Test

- Selection of f ratio:

  select ((ssb/dfs)/(ssw/dfw)) as f from details;

CHAPTER VI

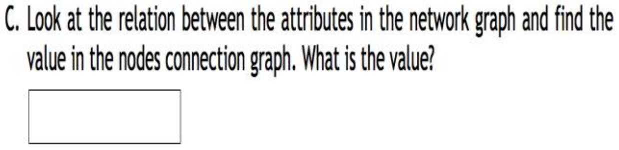
SECTION 5 (EVALUATION)

User study is conducted by evaluating the visualization in our approach giving tasks to users to test the understandability of the visualizations and the statistical tests. Various factors are considered and different interpretations and hypothesis are considered to test the interface. Section 5.1 describes the setup of the evaluation, section 5.2 describes the hypothesis considered, section 5.3 describes the data from the user study, section 5.4 explains the statistical analysis of the tests and section 5.5 gives the results of the hypothesis considered.

5.1 SET UP

Tasks were developed with various considerations. 21 tasks are given to each user, which have a combination of categories. 12 users are approached to perform the user study. These users were graduate students who had a minimum knowledge with statistical analysis and SQL. Tasks that are developed with time constraint for the calculation of efficiency are classified as efficiency based tasks. 4 such tasks are designed were each question is given to each test. A time constraint of 30 sec is given for each question. There are various complex questions, these include Anova and Friedman test questions and the less complex questions include chi square and t test questions. There are a total of 12 less complex questions and 8 complex questions. These are categorized as complex related tasks. One question is related to correlations. This is shown in the heat map. This task is classified under the visualization related tasks. 19 tasks has instructions where user select, observe and give the answers which as used as the learnability related tasks whereas 2 tasks has questions where visualizations are displayed to the user to answer the tasks related to these visualizations which are again related to visualization related tasks. Before the evaluation, every user had a 15 min of introduction session where an outline of the interface the functioning was explained in detail. A flowchart was presented to show the process flow of the tests. The initial 10 tasks are explained to the users simultaneous with the test to make the users understand and learn the processing. Total time is calculated for each user in minutes. Mean of the confidence ranging from 1 to 5 is calculated for each user. Properties like time, accuracy, confidence, complexity and efficiency are considered to evaluate the results of the user study to give the interpretation results. With the data from the user study, various hypothesis are considered and results are shown few using the statistical tests and few using the scatter plots. Accuracy, complexity and efficiency are taken as percentages whereas confidence is considered as mean. During the analysis of the complexity of data, tasks are divided into two groups. Group A contains the less complex tasks (chi square and t test) and Group B contains the complex tasks (anova and Friedman).

Table 5 shows the number of tasks under different categories of tasks developed and a example for each category.

Category of the tasks	**Number of tasks**	**Example**
Efficient based tasks	4	 C. Look at the relation between the attributes in the network graph and find the value in the nodes connection graph. What is the value?

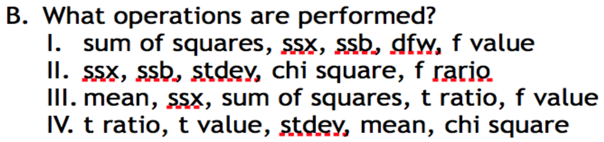
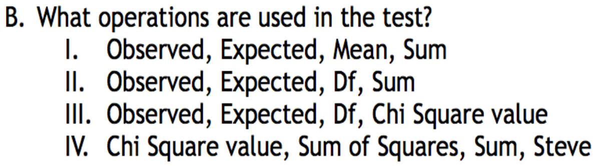
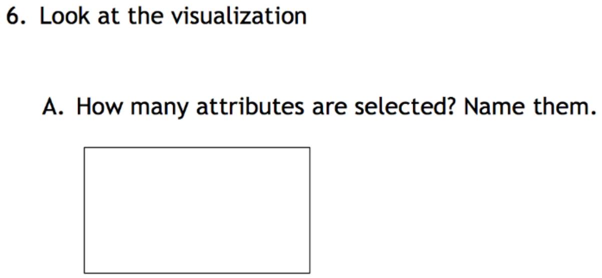
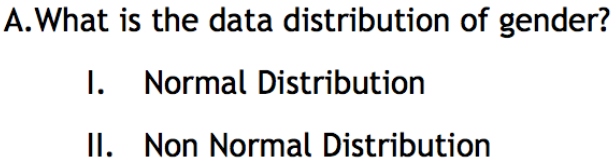
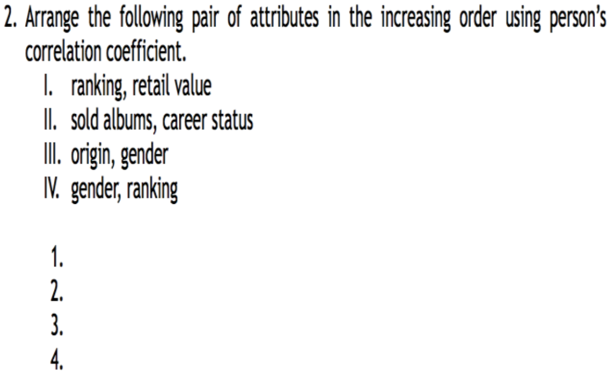
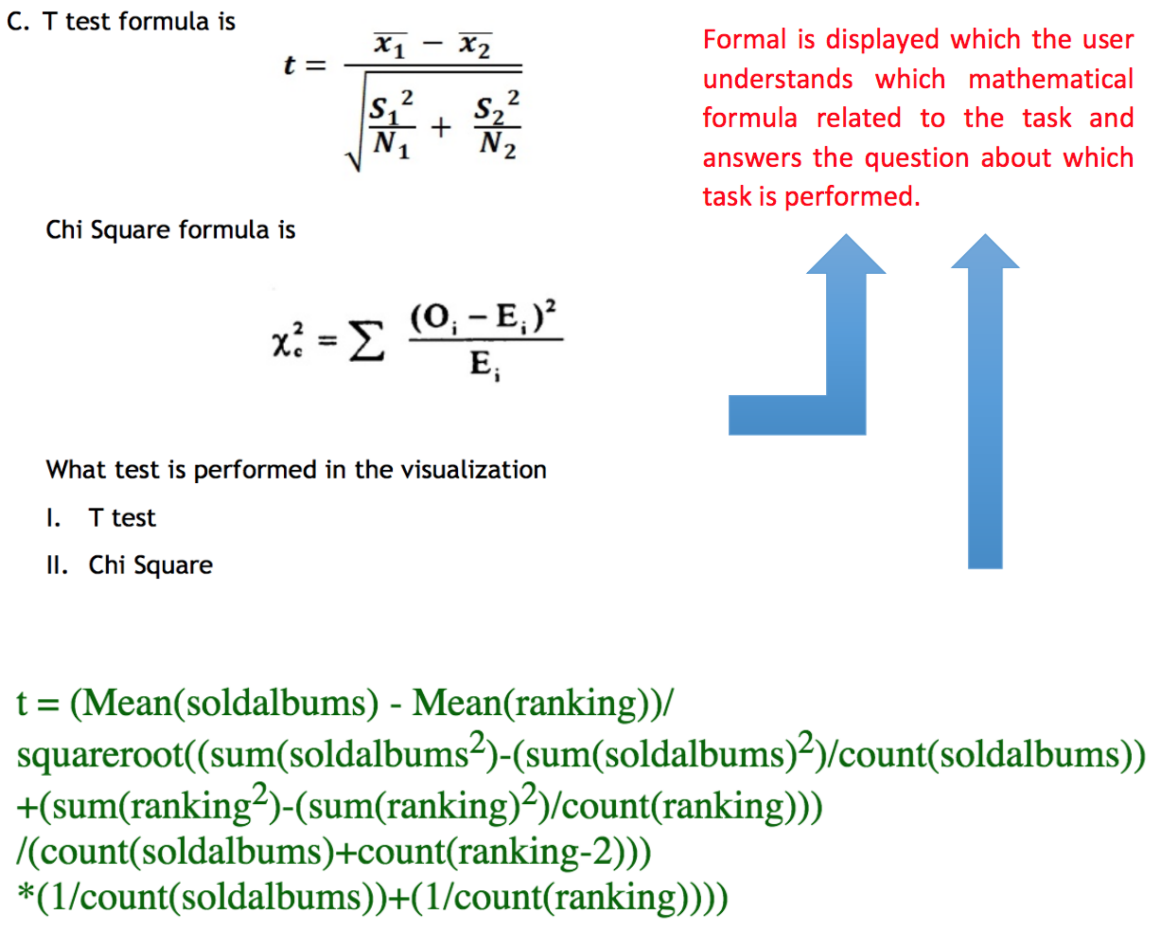
Complex tasks	8	 B. What operations are performed? I. sum of squares, ssx, ssb, dfw, f value II. ssx, ssb, stdev, chi square, f rario III. mean, ssx, sum of squares, t ratio, f value IV. t ratio, t value, stdev, mean, chi square
Less complex tasks	12	 B. What operations are used in the test? I. Observed, Expected, Mean, Sum II. Observed, Expected, Df, Sum III. Observed, Expected, Df, Chi Square value IV. Chi Square value, Sum of Squares, Sum, Steve
Visualization based tasks	2	 6. Look at the visualization A. How many attributes are selected? Name them.
Learnability based tasks	10	 A.What is the data distribution of gender? I. Normal Distribution II. Non Normal Distribution
Correlation tasks	1	 2. Arrange the following pair of attributes in the increasing order using person's correlation coefficient. I. ranking, retail value II. sold albums, career status III. origin, gender IV. gender, ranking 1. 2. 3. 4.

Table 1: Category of tasks and examples

Few examples are explained in details about the tasks developed. 2 categories are explained in detail. Fig 20 explains about the learnability based tasks.



C. T test formula is

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1{}^2}{N_1} + \frac{S_2{}^2}{N_2}}}$$

Chi Square formula is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

What test is performed in the visualization

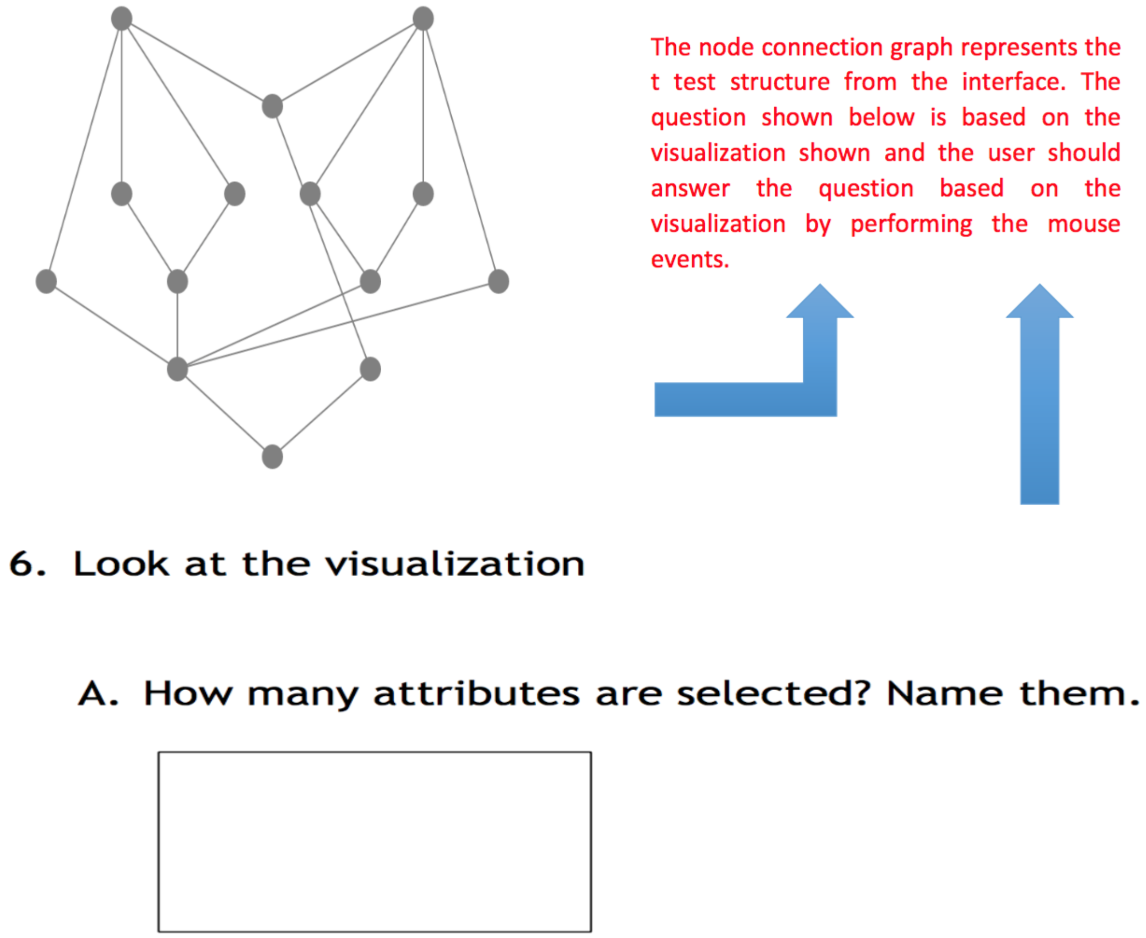
I. T test

II. Chi Square

Formal is displayed which the user understands which mathematical formula related to the task and answers the question about which task is performed.

t = (Mean(soldalbums) - Mean(ranking))/
squareroot((sum(soldalbums^2)-(sum(soldalbums)^2)/count(soldalbums))
+(sum(ranking^2)-(sum(ranking)^2)/count(ranking)))
/(count(soldalbums)+count(ranking-2)))
*(1/count(soldalbums))+(1/count(ranking))))

Figure 20: Learnability based tasks



The node connection graph represents the t test structure from the interface. The question shown below is based on the visualization shown and the user should answer the question based on the visualization by performing the mouse events.

6. Look at the visualization

A. How many attributes are selected? Name them.

Figure 21: Visualization based tasks

5.2 HYPOTHESIS

To interpret the results and the evaluation of the user study which further decides the understandability of the visualization and the statistical test methodologies. Different hypothesis considered are:

H1: Performance of user varies with the complexity of the visualization.

H2: Efficiency is time dependent.

H3: Accuracy and Confidence are directly proportional to each other.

H4: Time and Accuracy are inversely proportional to each other.

With the results of these interpretations we reject or accept the respective hypothesis. H1 and H2 are interpreted using the t-test calculation finding the statistical results whereas H3 and H4 are interpreted using the scatter plots.

5.3 DATA

Considering the properties taken in the setup, table 6 shows the data from the user study for the subject's time, accuracy, confidence and efficiency.

Users	**Time (mins)**	**Confidence (mean)**	**Efficiency (percentile)**	**Accuracy (percentile)**
1	21	4.04	100	100
2	27	3.61	50	90.4
3	32	4.23	100	100
4	18	4.23	100	100
5	34	4.33	75	95.2
6	20	4.42	75	95.2
7	18	3.80	50	90.2
8	24	3.90	50	90.2
9	30	4.42	100	100
10	19	4	75	95.2
11	16	4.04	100	100
12	36	3.33	50	90.4

Table 2: Data from the user study

Table 7 shows the data of the complexity of the tasks for group A and group B.

Users	**Group A**	**Group B**
1	100	100
2	100	75
3	100	100
4	100	100
5	100	87.5
6	100	87.5
7	83.33	100
8	91.66	87.5
9	100	100
10	100	87.5
11	100	100
12	100	75

Table 3: Data from user study for complexity

5.4 STATISTICAL ANALYSIS

Calculating the p value for H1 using the Wilcoxon test gives the result for the interpretation of the complexity and the accuracy hypothesis. Data from table 7 is considered for the calculation of the Wilcoxon test values. Since the data is not normally distributed Wilcoxon test is used else T test should be used for the statistical analysis.

- Group A: N = 12, Mean = 97.92
- Group B: N = 12, Mean = 91.67

Z value: -1.5213

P value: <0.05

The obtained p value < 0.05. The mean of group A and group B does not have the difference equal to zero.

Figure shows the histogram to represent the mean, square difference and the Square mean of the two groups.

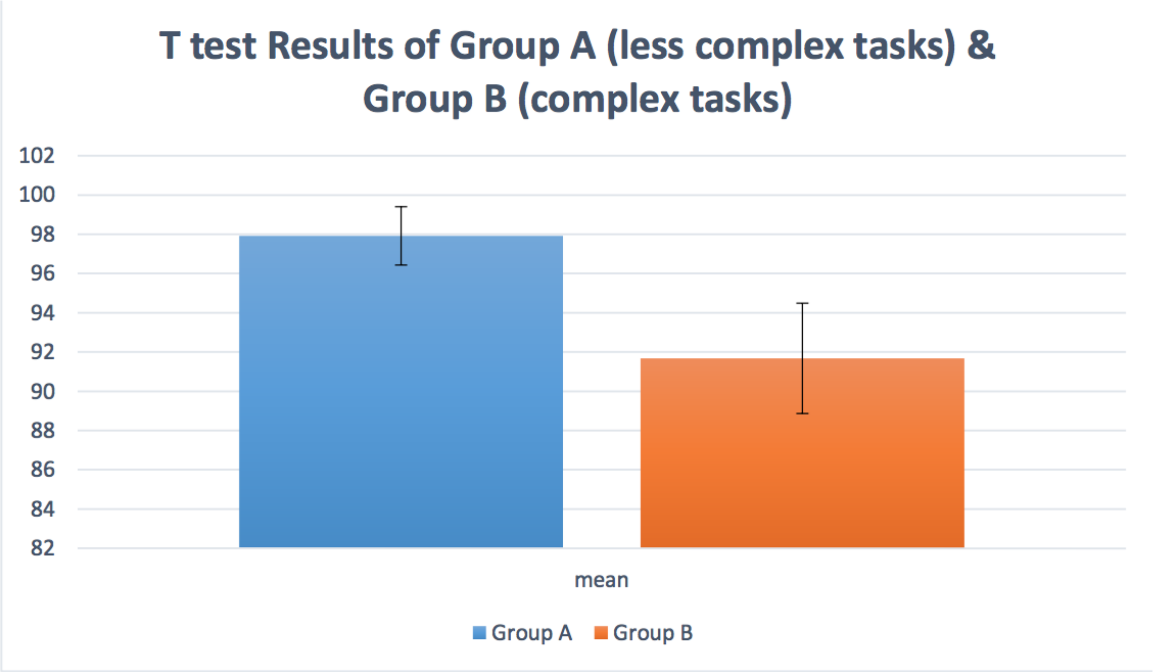


Figure 22: Wilcoxon test results of Group A and Group B

From figure 22 we can see that the values of Group A (chi square and t test tasks) less than Group B (anova and Friedman test tasks). The standard error bars are not overlapping which determines that the difference between both the means is not statistically significant. These give the idea of how precise the measurement is, or conversely, how far from the reported value is true.

5.5 RESULTS

Looking at the statistical results and the scatter plots we can conclude the interpreted hypothesis. Figure 23 shows the scatter plot for efficiency of the users, figure 24 shows the scatter plot for the accuracy and confidence of the users and figure 25 shows the time and accuracy.

Scatter plots in this case are used to interpret if the proposed hypothesis are accepted or rejected. Though the scatter plots do not exactly describe the interpretation results but the data and the clusters show the relation between each characteristic considered.

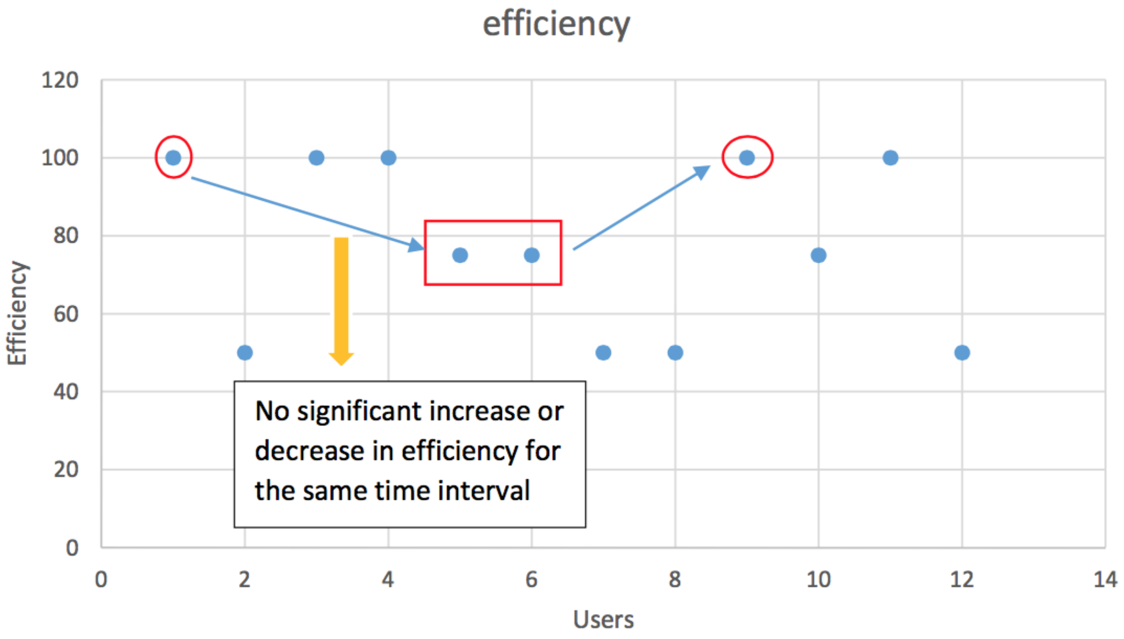


Figure 6: Efficiency of users with respect to time

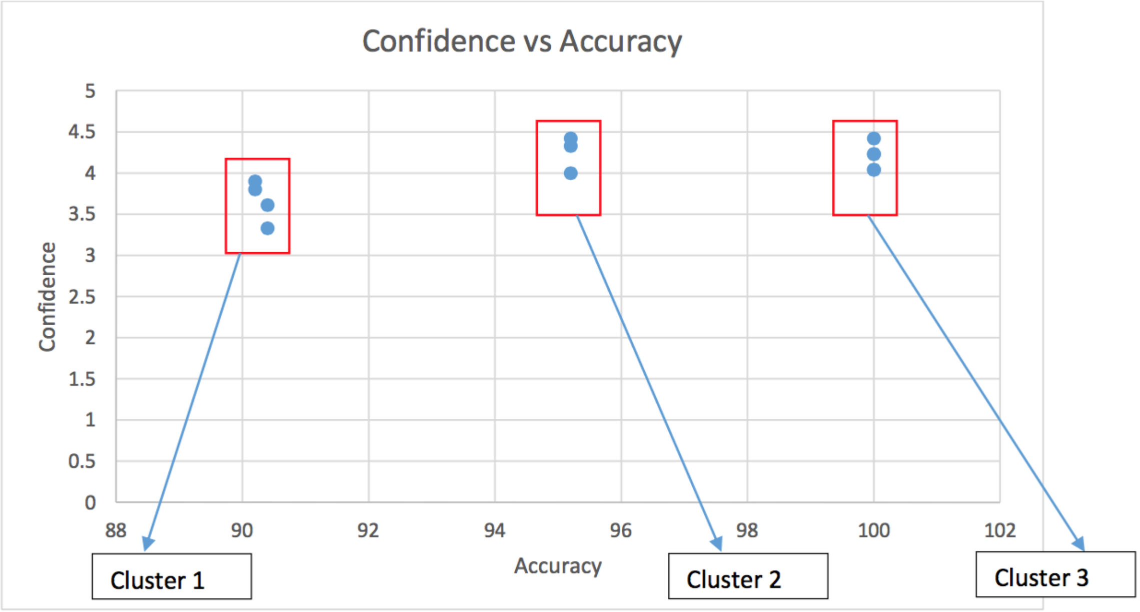


Figure 7: Confidence vs Accuracy

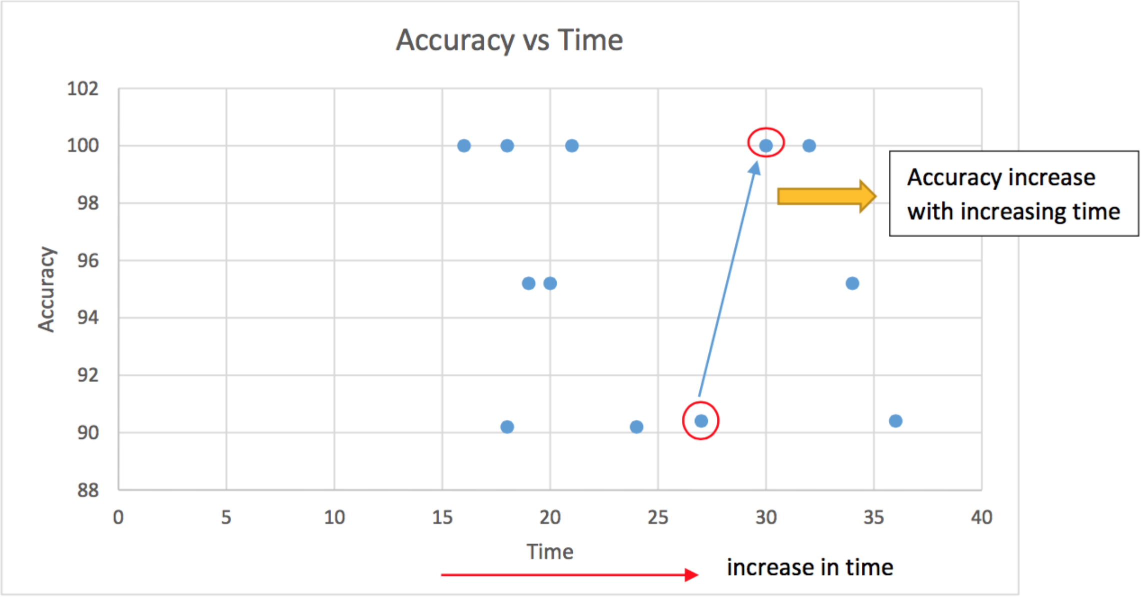


Figure 8: Accuracy vs Time

Results from the evaluation for the interpreted hypothesis:

H1: Statistical test is considered for this hypothesis. The Wilcoxon test results show a significant difference in the data. The obtained p value is less than 0.05. So we accept the hypothesis, which determines that the complexity of the tasks affect the accuracy. This indicates that the visualization with the anova and the t test are more complex to understand than the chi square and the t test.

H2: The scatter plot for the efficiency of the users is not dependent on the time factor. Figure 23 shows no clusters of data and the plots are segregated. These plots are not the same in the other scatter plots. Since we have a fixed time of 30 seconds for each question given to all the users, the data should not have a large variation. But we find many outliers, which do not match the other interpretations. But in the given 30 sec of time 41.6% of the users were able to answer these efficient related questions, which give us 100% efficiency. If the time limit can be increased to 40 or 45 sec the percentage of the users answering these questions will definitely increase which is positive to our assumed hypothesis. Thus efficiency plots are dependent on time. We can see an irregular increase and decrease of the data in the current plot, which can be varied with increase in the time. Thus we accept the hypothesis.

H3: Figure 24 shows the scatter plot for accuracy and confidence of the data from the user study. We observe three clusters, where every cluster varies its position. We can see a significant increase in accuracy with increase in confidence and vice versa. This determines that confidence and accuracy are dependent on each other. So we accept this hypothesis.

H4: Figure 25 shows the scatter plot for time vs accuracy. Though we see three linear clusters in the hypothesis, the outliers do not match the other scatter plots. The increase in time does not show any significant decrease in the accuracy for all the users consistently. In figure 23 we see the range of time is between 16 and 36. The accuracy does not have a regular increase or decrease in the data. For example, at time interval 27 we see the accuracy of 92% and at time interval 30 we see an accuracy of 100%. Our hypothesis interprets that there should be a decrease in the accuracy with increase in time or vice versa. The hypothesis is thus rejected, as there is no observed significance in the accuracy and time of the user study data.

CHAPTER VII

CONCLUSION AND FUTURE WORK

In this study, we proposed a web-based interface, which allows the users to select the attributes and visualize the statistical analysis of the data with SQL. Multi tabular data is used with nominal ordinal, categorical and continuous subjects. Categorical variables are ranked and views are created with these ranks replacing the original data. Query structures are visualized using nodes connecting graph. The node values of the trees are displayed on mouse over events. Formula for each node is displayed on click events. This makes the users easy to understand the query processing and test methodology. Network graph is constructed which shows the relation between the selected attributes. Depending on the type of data and the distribution of the data appropriate test is performed. In the network graph the thickness of the connection depends on the p value. Greater the thickness lesser the p value and this shows the relation between the pair of attributes. Correlations between every pair of attribute are shown as a heat map. Data distribution for each attribute is represented as a histogram with the normal curve and the difference is shown with the color variation. Evaluation is done on the visualization and the understandability of the visualizations with 12 users and the significant values are found for the time, complexity, correctness and confidence. With the results of our evaluation and the mentioned features, our interface can be used in many real world applications including the educational fields, by the data analyst for analysis of the data and for the statistical analysis of large databases. Developed visualizations in our approach can be used in online teaching of statistical test.

In our approach we used the nodes connecting graph and network graph for the visualization of the relation of the attributes and the queries. Better visualization techniques can be used for the network graph. As the number of attributes selection increases the complexity of the network increases. So better techniques can be used. The connections in the network graph are directly connected with the straight lines so this makes the graph ciaos with the increase in connections. So some of the connections can be replaced with the curves. So better visualization can be replaced. Due to the construction of the network graph with SVG circles and lines, due to the mouse events the highlighting of the connections and the circles are not displayed properly. So better visualization can be used to represent the attributes and their relation.

In the visualization we select the whole attribute and not a part of the attribute. We considered a musician dataset for the visualization, so if we need to select any two or more musicians and compare them. So in future we would like to write queries to select few rows and columns by the users.

In our approach we can only determine if the p value is < 0.05 or > 0.05, but the exact p value is not determined. In future we want to further extend our query processing methodology in finding the exact p value of the statistical tests.

Our approach can be used for any number of attributes but as the number of attributes increase the complexity to build the query for analysis increases. So queries can be modified and query processing can be improved.

Since our interface can be used for the learnability of the statistical tests and in teaching fields, the mouse events can be modified showing the values of the test results on the network graph and the node connection graph. This can improve the understandability of the tests and makes the learning process easier.

REFERENCES

[1] Domino: Extracting, Comparing, and Manipulating Subsets across Multiple Tabular Datasets, IEEE Transactions on Visualization and Computer Graphics (InfoVis '14), 20(12), pp. 2023-2032, 2014.

[2] Exploring Word-Sized Graphics for Visualizing Eye Tracking Data within Transcribed Experiment Recordings, ETVIS 2015.

[3] Databases will Visualize Queries too∗, http://www.vldb.org/.

[4] Computational Procedures for a One-Way ANOVA, http://www.graziano-raulin.com/tutorials/stat_comp/man1way.htm.

[5] Methods Manual: t-test - hand calculation - for independent samples, http://psychology.ucdavis.edu/sommerb/sommerdemo/stat_inf/tutorials/ttesthand.htm.

[6] The Friedman Test for 3 or More Correlated Samples, http://vassarstats.net/textbook/ch15a.html.

[7] Chi-Square Test for Independence, http://stattrek.com/chi-square-test/independence.aspx?Tutorial=AP.

[8] Using MySQL Data in JS Charts, http://www.zingchart.com/docs/tutorials/database-data/.

[9] Using Data Loader to connect charts to MySQL data base, http://www.amcharts.com/tutorials/using-data-loader-to-connect-charts-to-mysql-data-base/.

[10] Visualization-Aware Sampling for Very Large Databases, by Yongjoo Park.

[11] Exploring DATA Step Merges and PROC SQL Joins, SAS global form 2012.

[12] Computing the Pearson Correlation Coefficient, http://www.stat.wmich.edu/s216/book/node122.html.

[13] J2EE and MVC Architecture, Journal of Global Research Computer Science & Technology (JGRCST) Vol-I, Issue-II, July 2014.

[14] https://passel.unl.edu/pages/informationmodule.php?idinformationmodule=1130447119 &topicorder=14&maxto=15&minto=1.

[15] http://www.ehow.com/info_8647277_advantages-using-independent-group-ttest.html.

[16] imMens: Real-time Visual Querying of Big Data, Zhicheng Liu, Biye Jiang, Jeffrey Heer, Computer Graphics Forum (Proc. EuroVis), 32(3), 2013.

[17] NakeDB: Database Schema Visualization, Luis Miguel Cortes-Pena, Yi Han, Neil Pradhan, Romain Rigaux.

[18] Interactive visual summarization of multidimensional data published in Systems, Man and Cybernetics, 2009. SMC 2009.

[19] Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data [19] was published by Visualization and Computer Graphics, IEEE.

[20] Polaris: A System for Query, Analysis and Visualization of Multi-Dimensional Relational Databases [20] published by proceeding INFOVIS 2000.

[21] Perception-based Evaluation of Projection Methods for Multidimensional Data Visualization by Ronak Etemadpour, Robson Motta, Jose Gustavo de Souza Paiva, Rosane Minghim, Maria Cristina Ferreira de Oliveira, Member, IEEE, and Lars Linsen, Member, IEEE.

VITA

Vaishnavi Kamasani

Candidate for the Degree of

Master of Science

Thesis: VISUALIZING STATISTICAL ANALYSIS OF MULTI TABULAR ATTRIBUTES WITH SQL

Major Field: Computer science

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in May, 2016.

Completed the requirements for the Bachelor of Technology in Computer Science at GITAM University, Visakhapatnam, Andhra Pradesh, India in May, 2013.