Integrated case for visual analytics and statistical analysis using pension plan data

Kevin Pan
Samford University

Alan Blankley
Samford University

C. Clifton Eason
Samford University

ABSTRACT

This case integrates visual analytics and statistical analysis using real public pension data. You assume the role of an accountant working for the U.S. Department of Labor to evaluate public pension data to support decisions related to the public pension funding. This case emphasizes the use of critical thinking on numerical analysis of real data. Further, this case allows a student to conduct statistical analysis such as ANOVA (analysis of variance) using visual analytics while also understanding the concepts behind the analysis, not simply treating a software package as a black box. This case can be used in an undergraduate or graduate data analytics course, or can be used in an accounting, economics, finance, or marketing course. The competencies developed in this case are consistent with AACSB Accounting Standard A5.

Keywords: Data analytics, data visualization, statistical analysis, pension, accounting

INTRODUCTION

As businesses accumulate more and more data, there is a need for business schools to teach students data analytics. The importance of data analytics in business has been highlighted in numerous articles (Dadashzadeh, 2018; Kaufinger, 2019; LaMacchia, Rude, & Dolak 2017; Malaszczyk & Purcell, 2017; Nelson & Kazzaz, 2018; Park & Ellis, 2019; Purcell, 2013; Ragan, Chung, & Alibrandi, 2017; Ragan, Sparacio, McAndrew, & Filiciello, 2019).

In response to the need, AACSB recommends data analytics education in AACSB Standard 9 and 2018 AACSB Accounting Accreditation Standard A5 include reference to data analytics (AACSB, 2018). Two examples of data analytics in the AACSB Accounting Standard A5 are statistical techniques and visualization. While both statistical analysis and data visualization are becoming increasingly important for the business profession (Diamond & Mattia, 2017; Eaton, 2018; Hoelscher & Mortimer, 2018), they are not extensively covered in the conventional accounting curriculum. This paper provides an integrated case for business professors and students to learn visual analytics and statistical analysis using pension plan data.

This case is suitable for both undergraduate and graduate students. This case assumes that students have taken a basic introduction to data analytics course, but does not otherwise assume any prior experience in visual analytics or statistical analysis. The case can be used in a data analytics course, or can be used as part of an accounting, economics, finance, or marketing course. The case has 3 parts, and each part takes about 1 hour of class time.

For statistical analysis, this paper chooses to teach Analysis of Variance (ANOVA), an important tool in data analytics. Effectively a comparison of means among groups, an ANOVA is a helpful analysis in numerous business contexts: an advertising firm that wants to measure the differences in reactions to two versions of a commercial; a human resources executive who wants to assess the customer service ratings of multiple service teams; or an auditor evaluating the differences in claims payouts in different regions of the country. ANOVA has endless practical uses in multiple facets of business and, as such, is an important and relevant statistical tool for business students to learn. . The breadth of ANOVA's utility is one reason is serves as "the foundation of entire curricula in research methods courses in the social and behavioral sciences" (Gamst, Meyers, & Guarino, 2008, p.5).

Despite its utility and the prevalence of practical applications, the use of ANOVA in business curricula is often confined to statistics courses, and its practical value is sometimes under-emphasized in favor of ensuring students make the calculations correctly. The authors consider such situations a missed opportunity to bring the concept of ANOVA to a larger group of business students, teach its practicality in current business contexts, and emphasize its managerial usefulness. As such, the current authors use this manuscript to demonstrate that ANOVA can be easily taught outside of a statistics course and by faculty who are not statisticians, thereby benefitting more students.

For visual analytics, this paper illustrates the case in Tableau (Tableau Inc.), though it could also be conducted in other similar software programs. Tableau has been discussed in various articles in the literature (Aldihzer, 2017; Armadio & Haywood, 2019; Brands& Holtzblatt, 2015; Clayton, 2019; Enget, Saucedo, & Wright, 2017; Igou & Coe, 2016; Kokina, Pachamanova, & Corbett, 2017; Martin, 2016; Soule, 2018; Weirich, Tschakert, & Kozlowski, 2019;)

While visual analytics and statistical analysis can be learned via fictitious data, there is value in teaching by using real data (Best, 2019). Although the main purpose of this case is for students to learn visual analytics and statistical analysis, this case also exposes students to an important and interesting topic today: public pension crisis. This case does not require any prior knowledge on pension plans. The student does not need to have knowledge of pension plans to solve any of the problems. However, for students interested in learning more about the pension crisis, they can be referred to pension literature (Aubry & Crawford, 2017). Knowledge of the pension crisis could provide additional motivation for students to tackle this case.

CASE

Many private companies' pension plans are protected by the Pension Benefit Guaranty Corporation (PBGC). The PBGC was first established in 1974 when congress passed the Employee Retirement Income Security Act (ERISA) in an effort to provide an insurance program to protect retirees against the catastrophic loss of their retirement benefits (PBGC, 2019). The PBGC charges private companies premiums based on a flat fee for each plan participant plus a variable fee based on a measure of plan funded status. In this case, congress is considering establishing an entity similar to the PBGC to insure state and local pension plans against failure. In trying to determine appropriate insurance premiums for each state, the government's economists need to know whether there are regional differences associated with funded status, benefits paid, and contributions paid.

For purposes of this case, assume you are a second-year accounting staff member working for the Department of Labor in Washington, D.C. You have been appointed to a working team along with labor economists, lawyers, and actuaries from the PBGC to help assess the feasibility of insuring state pension plans. Your specific role is to analyze three measures of the "health" of state pension plans.

- Benefits paid to retirees – this is a measure of how much money each retiree, on average, gets paid from his or her pension;
- The required contribution as a percent of payroll – this ratio represents how much money the state or state agency contributes to the plan as a percent of total payroll;
- The funded status of the plan – this measure is frequently expressed as the ratio of plan assets to the plan liabilities. If assets equal or exceed the liabilities, the plan is considered fully funded or over funded. If assets are less than liabilities, the plan is considered to be underfunded. The lower the ratio, the worse the funding is at that point in time.

The team leader has asked you to calculate the national average for each of these variables as well as the average for each state and then rank results by state for each of the three variables. In addition, you are to determine whether there are differences in these measures across regions of the country.

After searching online for data, you find a data source that contains current and historical information about state public pension plans at the Center for Retirement Research at Boston College (available from http://publicplansdata.org/public-plans-database/download-full-data-set/).

Finally, after considering how best to analyze the data, you decide to conduct your analysis. These are your deliverables:

1. Develop visualization charts for funded status, benefits per retiree, and required contribution at plan level. Filter out null values and extreme values.
2. Find the national averages of funded status among all retirement plans.
3. Find the state averages of funded status. Sort the states in descending order by funded status. Develop visualization charts to illustrate the results.
4. Find the regional averages of funded status. Sort the regions in descending order by funded status.
5. Determine whether there are statistically significant differences in funded status across regions.
6. Write up an Executive Summary of your results supported by Tables, Charts, or Maps generated by Tableau.

REQUIREMENTS

Tableau software

Make sure Tableau software (www.tableau.com) is installed on the computer you'll be using. If you are a student or faculty member, you can download the Tableau software under the academic use license which is free to those using the software for academic purposes. It does take several days to gain approval, however, so if this step is needed, then allow at least a week prior to working this assignment to have the software installed.

Public pension plan data

When you are ready to begin the assignment, download the full dataset (plan level data) from the Center for Retirement Research at Boston College at the following address: https://publicplansdata.org/public-plans-database/download-full-data-set/
Choose "Download Full Data Set→" You will be downloading an Excel file named PPD_PlanLevel.xlsx. After downloading, store the file on the Desktop or in a directory of your choosing. Once the data are downloaded and you know where the file is, you are ready to begin.

Class Session 1

The learning objectives of this class session are data management, field calculation, and data filtering. You will also complete deliverable 1.

To import the data, open Tableau and then choose "Microsoft Excel" from the blue sidebar labeled "Connect." This tells Tableau that the file is an Excel spreadsheet. In the dialog box that opens, find the PPD_PlanLevel.xlsx file, click on it, then click Open. The sheet will open in the "Data Source" window. You are now ready to begin the analysis.

Next, you need to identify and create the fields needed to conduct the analysis. The dataset contains a large number of fields, but you will only use a few of them for this analysis. The public plans dataset contains most of the variables you want already. To evaluate the important details of the plans, you will use the variables shown in Exhibit 1.

You should look up the variable definitions in the Data Dictionary, because it is important to understand what each variable represents before including or excluding it from the analysis. The data dictionary for the Public Plans dataset can be found at: https://publicplansdata.org/public-plans-database/documentation/

To create a calculated field, in the Data Source window, click on the drop-down arrow of any field and select "Create Calculated Field." It will be more convenient if you start by selecting a field that will be part of your calculation, but it is not necessary. The name of the field you clicked on will be in braces [Field_Name] in the Create Field dialog box. If it is not a part of the calculation, then delete it and type in the two relevant fields. Tableau will assist by bringing up a list of fields as you begin entering the names of the fields. You should create a calculated field for funded status, which will be the ratio of the market value of plan assets to the plan liabilities. In the Create Field box, key in [MktAssets net]/[ActLiabilities GASB]. Name the new calculated ratio field, FundedRatio_Mkt.

To rename a field, in the Data Source sheet, scroll over to find the field you wish to rename. Click the drop-down arrow in the upper right corner of the field name, then select rename. Type the new name into the name box. When finished, hit return. You will rename the field, ReqContRate_ER, "EmpContribution."

In this analysis, you will only use data from 2017. You can relax this assumption if desired. To restrict the analysis to data from 2017, you will create a Filter. In the Data Source window, click on "Add" under "Filters" in the top right of the screen. Click "Add" in the dialog box. The variable for Fiscal Year is FY, so scroll down and select "FY" from the list. Exhibit 2 shows the filter range window in Tableau. Enter range of values as 2017 to 2017. After clicking OK, the dialog box, as shown in Exhibit 3, then tells you that it will keep 2017, and by extension filter out the other years. Click OK. Tableau will apply this fy filter to the entire project.

Next, you can work on Deliverable 1: Develop visualization charts for funded status, benefits per retiree, and required contribution at plan level. Filter out null values and extreme values.

To develop visualization charts, click on Sheet 1. When you use Tableau, it is a good habit to always name a worksheet according to your objective for this worksheet. You can simply double click on "Sheet 1" to rename it. Rename this worksheet "Plan Level Analysis." The worksheet interface in shown in Exhibit 4.

In Tableau, the analysis of the data happens in worksheets. In a worksheet, a field is either a dimension or a measure: dimensions are qualitative fields and measures are quantitative fields. Dimensions and measures are listed on the left side of the window. For example, "State" is a dimension. The fields you are interested in analyzing – funded status, benefits per retiree, and required contribution – are measures.

When you create a chart or a table, Tableau will always summarize a measure by a statistic such as sum, average, or variance by default. If, as in this case, you merely want to see the value of a particular field for each record, then each record must be unique. In this case, there are 162 unique public plans in 2017 listed in the dataset. Since Plan Name is the value for each row, Tableau finds the value for each corresponding FRatio_Mkt, and then will sum all the values of FRatio_Mkt for that particular Plan Name. Since there is only one value of FRatio_Mkt per Plan Name, the sum of a number is always the number itself, as is the average of that number. For this reason, when each row in a chart has only a single value of a variable associated with the row, it does not matter whether Tableau shows a sum or average in the column.

To create a chart, you can simply drag fields from Dimensions or Measures to Columns or Rows. For this deliverable, you want to have a horizontal bar chart with each plan listed in the dataset as a row, and the bar chart shows each plan's funded status. To do so, you simply need to drag Plan Name from Dimensions to Rows, and then drag FundedRatio_Mkt to Columns. To find a specific field by name, you can enter text in the Search box above Dimensions. For example, to find the created field FundedRatio_Mkt, enter FundedRatio_Mkt, then it will appear in the Measures window. You can then drag it to columns. Note that, by default, Tableau displays the FundedRatio_Mkt variable as "SUM(FundedRatio_Mkt)." You can change this, as you did, to "AVG(FundedRatio_Mkt)," but in this case the particular statistic selected does not change the result since there is only one value for each plan in the data set.

You can sort a chart by clicking on the Sort button in the tool bar. Once you sort it, you will see a sorted bar chart. If you scroll to the bottom of the chart, you will see the following chart that says "16 nulls" beginning with the Utah Noncontributory plan and followed by 15 others all showing nulls. This is shown in Exhibit 5.

To filter out the null value in FundedRatio_Mkt, click on Data Source window, then click on "Add Filters." Add a filter to FundedRatio_Mkt. The filter interface is shown in Exhibit 6. Click OK; note that the Include Null Values box is not checked. Click OK again in the next dialog box. There are still nulls in the data set, but they are not in the variables you want to analyze. When you click on the Plan Level Analysis sheet, you will see that the Utah Noncontributory plan as well as the other nulls have now been filtered out.

Now repeat the above steps for the AvgBenefit variable. Exhibit 7 shows the results. Note that the plan with the maximum value, the Utah Public Safety Plan, displays a value of close to 32,000 (actually 32 million – the numbers are in thousands), while the next highest value is the Miami Fire and Police plan, which displays a value of less than 200. Clearly, the Utah Public Safety Plan is an outlier that distorts the scale of the chart so that the visualization of the results is much less informative than desired. In this case, you will simply filter this observation out. Go back to the Data Source sheet and Click "Edit Filter." Click on "Add" and then search for AvgBenefit. Exhibit 8 shows the resulting dialog box. In the box displaying the max value, change the number from 31,749 to 200, as in Exhibit 9. Then click OK.

Returning to the Plan Level Analysis sheet will reveal that the chart has automatically changed to reflect the new, filtered results. Exhibit 10 displays the results.

Class session 2

The learning objectives of class session 2 are statistical and visual analysis at different levels, and data joining. You will complete Deliverables 2-4.

This class session teaches students how to analyze data visually and statistically at different levels of interest, namely, national level, state level, and regional level in this case. Although you focus exclusively on the funding ratio, you could also do the same thing for the other two variables of interest as well following the same steps.

Deliverable 2 is to find the national averages of funded status for all retirement plans. To do so, create a new worksheet. Name it Averages. Drag FundedRatio_Mkt to the "Text" box located under the "Marks" panel, as shown in Exhibit 11.

You will need to change Sum to Average to find the average, which is 0.7340. To change Sum to Average, click the drop down arrow on the Sum(FundedRatio_Mkt) field located in the Marks panel will reveal the "Measure" choice, which will allow you to select Average as in the menu structure in Exhibit 12.

Deliverable 3 is to find the state averages of funded status; sort the states in descending order by funded status; develop visualization charts to illustrate the results.

As part of the analysis, you will want to make comparisons of plan funded status by region and state. To find the state averages, you can simply drag State Name from Dimensions to Rows. Tableau will now have each state's average funded status. Visualizing this in a bar chart only requires clicking on the bar chart icon in the Show Me tab. The result is shown in Exhibit 13. Sorting by descending order shows that District of Columbia has the most well-funded pension plan of any state or district at 1.0659. South Dakota has the most well-funded pension obligations among the states at 1.0008. Both plans have assets (at fair value) exceeding the liabilities of the plans. On the other hand, Illinois has the lowest average funded status of all the states at 0.4760.

To view the same results on a map of the United States, you need to add a "geographic role" to the State Name field. Exhibit 14 demonstrates how to change the State Name from a text field to a Geographic Role field. Select State Name, click on the drop-down arrow next to State Name and follow the menu to Geographic Role/State/Province.

After you add the geographic role, you can simply click on the second map icon in the second row of the Show Me tab. Results will appear as in Exhibit 15.

Deliverable 4 is to find the regional averages of funded status; sort the regions in descending order by funded status.

The previous state-level analysis shows visually that there might be geographical variation across the regions of the United States: Northeast, South, Midwest, West. Therefore, it would be interesting to quantify the geographical variation across the regions. Since there is no "Region" variable in the original dataset, you need to join a second file to the original. The purpose here is to expose students to this very common data management need and to show how easy this is to do in Tableau.

While there may be other ways to group the regions in the United States, for simplicity you group the states into four regions as shown in Exhibit 16. Copy Exhibit 16, a list of State abbreviations with the associated region, into Excel and create a new file named Region.xlsx.

To join the new Excel file with the data already in tableau, there needs to be a common field between the two datasets. In this example, the State Abbreviation field in each file is the common field. The two fields do not need to be named the same name for them to be joined.

To join the Excel file containing the State and Region, go to the Data Source page and click on "Add" to the right of the Connections list. Select Microsoft Excel. In the dialog box, find the Region.xlsx file in the directory where you stored it, and select the file name. Click Open. The join box, as shown in Exhibit 17, will appear.

At the drop-down arrow under Data Source, select the State Abbrev field and join it the State field in the Region.xlsx file. You want the inner join. A discussion of different types of joins in beyond the scope of this case. You can check to see if it is loaded correctly by scrolling across the data source page to the far right. Close the join dialog box.

To find each region's average funded status, return to the bar chart of state averages. You can simply replace State in Rows by Region in Rows. After you sort the bar chart, you will have the bar chart shown in Exhibit 18. It shows that the South has the highest funded status at 0.7688, and the Northeast has the lowest funded status at 0.6505.

Class session 3.

The learning objective of class session 3 is ANOVA (Analysis Of Variance). You will also complete deliverable 5.

This class session teaches students how to evaluate statistical significance of variation across groups; in this case, you will evaluate whether the regions' average funded ratios differ statistically from one another. This case integrates visual analytics and statistical analysis, as illustrated below.

Deliverable 5 is to determine whether there are statistically significant differences in funded status across regions.

Since there is a visually observable difference in the average funded ratios across the regions of almost 12 percentage points, it appears that the average funded status differs across regions, but you do not know whether the difference that you observe across the regions is statistically significant. Being statistically significant means that the observed difference did not occur by chance; it is unlikely that the difference came about by random chance. To test whether the difference is statistically significant, you should use Analysis of Variance or ANOVA. ANOVA is a hypothesis testing method used to determine whether there is a statistically significant difference in a variable across groups. In this case, you want to know whether the average funding ratio differs across the group of geographic regions.

The main idea of ANOVA is to compare the between-group variation to the within-group variation. The within-group variation offers a baseline level of variation against which you can compare the variation between the groups, or regions. If the between-group variation is large compared to the within-group variation, then the between-group variation is statistically significant.

To assess if there is statistical significance, it is necessary to set up a null hypothesis and the alternative hypothesis correctly. Let's start with the alternative hypothesis (H_a), which is the hypothesis that you wish to establish. In ANOVA, the alternative hypothesis is that there is significant variation between groups. Mathematically, this means that at least one group has a different mean from the other groups. In this case, the alternative hypothesis is that at least one region's average funded status is different from the average funded status of the other regions. The null hypothesis (H_0) is the opposite of the alternative hypothesis; you wish to reject the null hypothesis so that it will be unlikely that the alternative occurs by chance. H_0 in ANOVA is that there is no significant variation between groups. Mathematically, H_0 states that all groups have the same mean. In this case, the null hypothesis is that all regions have the same average funded status.

An advantage of ANOVA is that it compares all the groups at the same time. It therefore offers better control of type I error, i.e., falsely rejecting the null hypothesis, than comparing two groups at a time. If the p-value, the probability of type I error, is less than "α," you can reject the null hypothesis. "α" is the significance level, the largest acceptable probability of a false conclusion. α is predefined before the computation. A common choice of α is 0.05.

In ANOVA calculations, the goal is to find the "F" statistic; once "F" is found, the p-value can be estimated. "F" represents the ratio of between-group variation to within-group variation (F = Between-group variation/Within-group variation). In order to determine whether the differences you observed in the funded ratio across regions, you can use Tableau to calculate the variation, and then can perform the arithmetic in a series of simple steps to determine the F-statistic and p-values.

ANOVA step 1: calculate sum of squares (SS)

"Sum of Squares" is the measure statisticians use to determine variability. In general, this measure represents the difference between each observation and the average. Because some observations are less than the average, and some are greater than the average, you square the difference for each measure, and then sum these squared differences to determine variability. Since you want to compare the between-group variation to the within-group variation, you need to start by finding the total variation for all the data, without regard to the groups, then find the within-group variation, and finally subtract the within-group variation from the total variation. In fact, calculating these variation amounts is usually the hard part of using ANOVA, and statisticians enjoy explaining the mathematics underlying this effort with very impressive looking formulas. Rather than calculating the different Sums of Squares directly, however, you will use Tableau to help us calculate them indirectly. Tableau can generate a "Variance" statistic for any variable. "Variance" is a very specific measure of variability that statisticians use frequently, but for your purposes, it includes, as part of the mathematics, the sum of squares. In the following formulas, you can "back into" the sum of squares from the variance statistics produced by Tableau as follows:

You can find the "Total Sum of Squares" (SS_Total) as follows:

SS_total = Total variance * (total sample size - 1)

It is important to understand that total variation is equal to the within-group variation plus the between-group variation, since these are the only two forms of variation that can occur in the data. Mathematically, SS_total = SS_within + SS_between.

We can now determine within-group variation as follows:

SS_within = Sum(group variance * (group size -1))

Since we now know both total and within-group variation, you can determine between-group variation as the difference between the two as follows:

SS_between = SS_total – SS_within

To find the sum of squares you need to perform an ANOVA, start by using Tableau to find the variance of the FundedRatio_Mkt variable.

First, search for the Number of Records variable; Tableau automatically creates this variable and lists it under Measures. Then drag this variable to the Text box on the Marks panel. Then drag the FundedRatio_Mkt variable to the Text box as well. Change the measure from Sum to Variance.

The results are shown in Exhibit 19. There are 159 records in the dataset, with a variance of .03097. This information allows us to calculate the SS_total, which is

SS_total = Total variance * (total sample size - 1)
= 0.03097 * (159-1) = 4.89326

Next, to find the variance of each group, you can simply drag Region to Columns, and Tableau will automatically calculate the variance of each group. This is one of the most useful features of Tableau, as it can allow a user to find a measure at different levels. A key learning objective of this case is to show that a student can visualize a complex statistical procedure such as ANOVA through the use of the Tableau interface. The results are shown in Exhibit 20.

Once you have the number of records and the variance for each region, you can easily calculate the within-group sum of squares, SS_Within as follows:

SS_within = Sum(group variance * (group size -1))
= [0.03475 * (43-1)] + [0.03896 * (20-1)] + [0.02937 * (55-1)] + [0.02083 * (41-1)]
= 4.61892

To find the between-group variation, you can now calculate the between-group sum of squares as:
SS_between = SS_total – SS_within
= 4.89326 – 4.61892
= 0.27434

ANOVA step 2: calculate mean squares (MS)

In order to ultimately calculate the F statistic, you have to first find what are referred to as "Mean Square." In general, "Mean Square" is an estimate of the average variation between or within groups, depending on how it's calculated. Like the sum of squares, you want to determine both within-group and between-group Mean Square values.

To determine between-group Mean Squares, you will use the following formula:
MS_between = SS_between / (number of groups -1)

To determine between-group Mean Squares, you will use the following formula:
MS_within = SS_within / (total sample size – number of groups)
In this case,
MS_between = SS_between / (number of groups -1) = 0.27434 / (4-1) = 0.0914467

MS_within = SS_within / (total sample size – number of groups) = 4.61892 / (160-4) = 0.0297995

ANOVA step 3: calculate the F-statistic and p-value

Once you know the two Mean Square values, the F-statistic can easily be calculated by dividing MS_between by MS_within as follows:
F = MS_between/MS_within
= 3.0687

In the final step, you need to determine whether the F-statistic is statistically significant or not. If it is, you can say with confidence that the FundedRatio_MKT differs across regions. In order to determine whether an F-statistic with a value of 3.0687 is significant, you can look up the F-statistic value in an F-distribution critical value table (which can be found in many statistics textbooks and many Web sites, e.g., http://www.socr.ucla.edu/applets.dir/f_table.html), with numerator degrees of freedom = 4-1=3 and denominator degrees of freedom = 156-1 = 155. Depending on the table you look at, the denominator degrees of freedom may only show 120 or infinity. Using 120 (rather than 155, as in this case) will be a more conservative estimate, and shows the critical F-statistic value is 2.68. Since your calculated F value is greater than the F_critical value of 2.68, the p-value is less than 0.05. Since the p-value is less than α, you can reject the null hypothesis; i.e., the regional variation is statistically significant.

Your final homework assignment is to complete Deliverable 6, write up an Executive summary of your results supported by Tables, Charts, or Maps generated by tableau. You should show the Tableau results and calculations you used to do the Analysis of Variance test.

TEACHING NOTES

Learning objectives of this case:

* Critical thinking: the students think critically about the range of data.
* Data Management (data filtering, data joining)
* Visual analytics: not only create charts, but learn to use a visual interface to create analytics
* Statistical analysis: e.g., hypothesis testing with ANOVA

Student learning outcomes:

At the completion of this case, the student should be able to:
* Use Tableau to summarize data by charts, tables, and maps
* Perform data filtering to relevant fields
* Perform data joining to combine data from different sources
* Examine data at different levels (e.g., national, state, region) via visual analytics
* Understand the concept of hypothesis testing and ANOVA
* Perform a simple ANOVA analysis

One of the questions that students might ask is, "Why do I have to learn how ANOVA is calculated? Can I just not click on a button on a software package to do that in 2 seconds?" To be able to apply software correctly, a business professional must understand what a statistical calculation is and how it works. Otherwise, a business professional can incorrectly apply a method to a problem, resulting in misleading business conclusions and material loss for a business and its stakeholders. Because of its practical relevance in the workplace, students with a firm grasp of the principles of ANOVA will have a great asset to bring to any job that requires analysis. Further, being able to perform an ANOVA and interpret its results using visual analytics could be a skill that makes ANOVA a more accessible approach for smaller businesses and those that have not invested in expensive analytics software.

And while many business students and some business professionals do have access to expensive software tools such as SPSS and SAS, which can make performing an ANOVA quite easy, a number of limitations still exist:

1) Many in industry do not have access to high-end analytics software, which can be cost-prohibitive to many firms. While this is particularly true for smaller firms that may lack the funds or personnel to effectively use these programs, it is also the case for many larger companies that do not engage in data analysis frequently enough to justify the expense.

2) Big data is becoming increasingly voluminous. Some analytics software does not handle large amounts of data efficiently.

3) Businesses often maintain their data in a native format that is not directly compatible with higher-end analytics software. In such cases, data files must be converted to a format that is acceptable for the analytics program. The process of such a conversion can be difficult and time-consuming.

4) Perhaps most importantly from an educational standpoint, when a student/user clicks a button to perform an ANOVA in analytics software, many users lack an understanding of the calculations being formed "behind the scenes" and may not truly apply critical thinking or logical judgment in interpreting the results or considering business implications or managerial actions that may result from the output.

In summary, many business users do not have access to expensive analytics software. Those who do may spend more time converting their data into an acceptable format than it would take to simply perform an ANOVA in Tableau. Further, the ease of clicking a "run" button shields the user – whether a professional or a student – from understanding the components that influence the output. As such, business educators may find that teaching ANOVA using Tableau addresses the above limitations and provides students with a valuable skill undergirded by a true understanding of both the process and the relevance of ANOVA.

Optional exercise: incorporate Excel

You can also use an Excel function to estimate the p-value instead of using a table. Alternatively, you could use Microsoft Excel to find the p-value, and the following formula can be used:
p-value = F.DIST.RT(F, number of groups -1, total sample size – number of groups)
In this case, the estimated p-value using this function in Excel is
=F.DIST.RT(3.0687, 3, 156) = 0.02965, which is less than .05.

Optional exercise: find group variances using SQL

Variance can be found in database management systems that interface with big data. In SQL, SELECT VAR allows a user to find variance. This will allow you to find total variance. Then, you can use SLECT VAR GROUP BY to find group variance.

Optional exercise: What to do next if you could reject the null hypothesis?

If you reject the null hypothesis, you could further analyze the data by performing post-hoc tests. After testing a hypothesis using ANOVA, people often use analytics software programs to perform post-hoc tests. In this case, you might want to know, which regions differ from one another. You have already seen that there are differences across the regions, but you do not know whether, the South differ from, say, the West region. Post hoc tests use variances as above, so you could also simply calculate them using the numbers you have. For example, one post-hoc test, the Bonferroni test, can be computed as follows:

$$\alpha_{\mathrm{corrected}} = \frac{\alpha}{\text{number of groups} - 1}$$

$$t = \frac{\overline{x_i} - \overline{x_j}}{\sqrt{\mathrm{MS}_{\mathrm{within}}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

For example, to compare the Northeast region with the South region, you would calculate the $\alpha_{\mathrm{corrected}}$ as .05/3 =0.0167. Then you would calculate t (this is a t-statistic for a t-test of the means of the two groups involved) as (.7688 – .6505)/ (SqRt of .0297995(1/55 + 1/20)) = 2.6245. Using Excel to calculate the p-value indicates that the p-value is equal to .0105.

Comparing that value to the corrected alpha of .0167, indicates that the South average FundedRatio_Mkt is significantly differ from the Northeast average FundedRatio_Mkt. There are various post-hoc tests, each with their pros and cons. For example, the Bonferroni test is more complicated but can deal with unequal group sizes. A simpler test that deals with equal group size is Tukey's test. A detailed discussion of various post hoc tests is beyond the scope of this article.

REFERENCES

AACSB Standards Web site https://www.aacsb.edu/accreditation/standards/accounting. Retrieved February 5, 2019.

Aldhizer III, G. R. (2017). Visual and Text Analytics: The Next Step in Forensic Auditing and Accounting. *CPA Journal*, 87(6).

Amadio, W. J., & Haywood, M. E. (2019). Data Analytics and the Cash Collections Process: An Adaptable Case Employing Excel and Tableau. In *Advances in Accounting Education: Teaching and Curriculum Innovations* (pp. 45-70). Emerald Publishing Limited.

Aubry, J. P., & Crawford, C. V. (2017). State and Local Pension Reform Since the Financial Crisis. *Center for Retirement Research at Boston College: State and Local Pension Plans*, *54*.

Best R. W. (2019) Teaching bank profit decomposition using real-world data. *Journal of Business Cases and Applications*, 23.

Brands, K., & Holtzblatt, M. (2015). Business Analytics: Transforming the Role of Management Accountants. *Management Accounting Quarterly*, 16(3).

Clayton, P. R., & Clopton, J. (2019). Business curriculum redesign: Integrating data analytics. *Journal of Education for Business*, *94*(1), 57-63.

Dadashzadeh, M. A case study to introduce Microsoft Data Mining in the database course. (2018). *Journal of Business Cases and Applications*, 22.

Diamond, M., & Mattia, A. (2017). Data Visualization: An Exploratory Study into the Software Tools Used by Businesses. *Journal of Instructional Pedagogies*, 18.

Eaton, T. V., & Baader, M. (2018). Data Visualization Software: An Introduction to Tableau for CPAs. *The CPA Journal*, 88(6), 50-53.

Enget, K., Saucedo, G. D., & Wright, N. S. (2017). Mystery, Inc.: A Big Data case. *Journal of Accounting Education*, 38, 9-22.

Gamst, G., Meyers, L. S., & Guarino, A. (2008). Analysis of variance designs: A conceptual and computational approach with SPSS and SAS. New York, NY: Cambridge University Press.

Hoelscher, J., & Mortimer, A. (2018). Using Tableau to visualize data and drive decision-making. *Journal of Accounting Education*, *44*, 49-59.

Igou, A., & Coe, M. (2016). Vistabeans coffee shop data analytics teaching case. *Journal of Accounting Education*, 36, 75-86.
Kaufinger G. G. (2019) Beautiful Homes Inc.: A Microsoft Excel case for a business computing applications course. *Journal of Business Cases and Applications*, 23.

Kokina, J., Pachamanova, D., & Corbett, A. (2017). The role of data visualization and analytics in performance management: Guiding entrepreneurial growth decisions. *Journal of Accounting Education*, 38, 50-62.

LaMacchia, C., Rude J., & Dolak, E. (2017). Big data decision making: an application activity. *Journal of Business Cases and Applications*, 17.

Malaszczyk, K., & Purcell, B. M. (2017). BIG DATA ANALYTICS IN TAX FRAUD DETECTION. Northeastern Association of Business, *Economics and Technology*, 233.

Martin, F., & Ndoye, A. (2016). Using learning analytics to assess student learning in online courses. *Journal of University Teaching & Learning Practice*, 13(3), 7.

Nelson M. L., & Kazzaz M. (2018) A case of the subrogation blues! A business analyst's sourcing recommendation. *Journal of Business Cases and Applications*, 20.

Park T., & Ellis Y. (2019). The case of Grizzly Sports Highlighted, Inc.: analyzing accounting data using Excel. *Journal of Business Cases and Applications*, 23.

Pension Benefit Guaranty Corporation (PBGC) Web site https://www.pbgc.gov/. Retrieved February 5, 2019.

Purcell, B. (2013). The emergence of" big data" technology and analytics. *Journal of Technology research*, 4, 1.

Ragan J. M., Chung D., & Alibrandi V. (2017). STARExplorer Case: Searching for ways to integrate data analytics and accounting. *Journal of Business Cases and Applications*, 18.

Ragan J. M., Sparacio G. P., McAndrew C. P., & Filiciello A. J. (2019). The Evolving Global Enterprise: Preparing Future Accountants Using Analytics and Systems Integration. *Journal of Business Cases and Applications*, 23.

Soule, L., Fanguy, R., Kleen, B., Giguette, R., & Rodrigue, M. S. (2018). EVOLUTION OF A FIRST COURSE IN DATA ANALYTICS FOR BUSINESS STUDENTS. *Journal of Research in Business Information Systems*, 2014, 55.

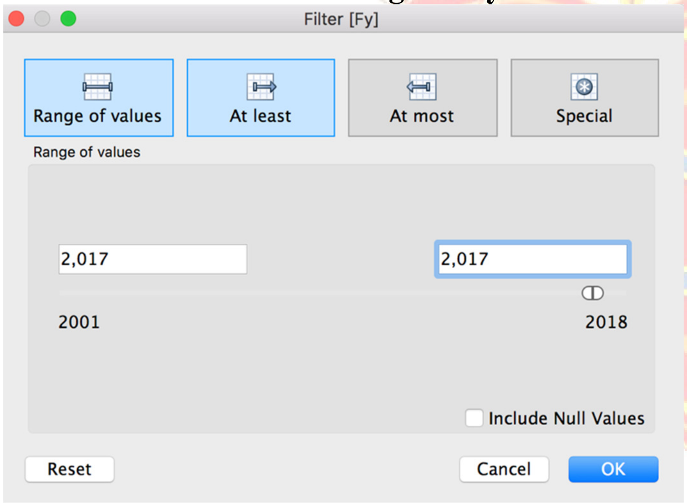
Weirich, T. R., Tschakert, N., & Kozlowski, S. (2019). Teaching Data Analytics Skills in Auditing Classes using Tableau. *Journal of Emerging Technologies in Accounting*.

APPENDIX

Exhibit 1. Variables to analyze in this case

Concept Analyzed	Variable in Dataset
Benefits paid to retirees	BeneficiaryBenefit avg ($000s) (You will rename this field to AvgBenefit).
Required Contribution percent of Payroll	ReqContRate_ER (You will rename this field to EmpContribution)
Funded Status	MktAssets_net/ActLiabilities_GASB (You will name this calculated field FundedRatio_Mkt)

Exhibit 2. Tableau filter range for fy



Filter [Fy]

Range of values

2,017

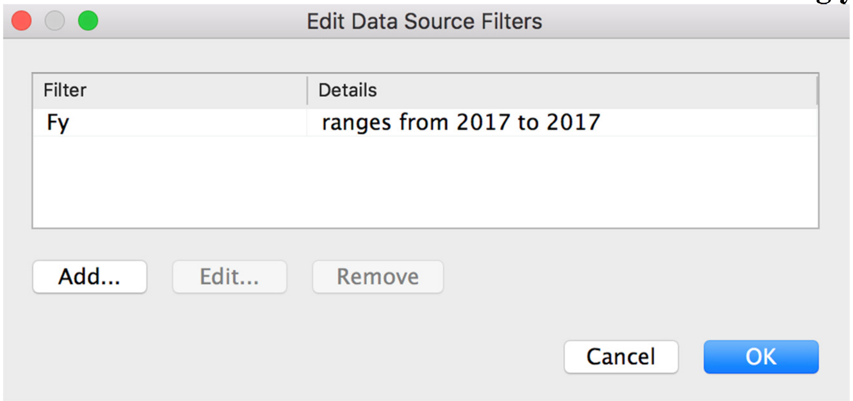
2,017

2001

2018

□ Include Null Values

Exhibit 3. Tableau Edit Data Source Filters window showing you have filtered for 2017



Edit Data Source Filters

Filter	Details
Fy	ranges from 2017 to 2017

Exhibit 4. Tableau worksheet "Plan Level Analysis."

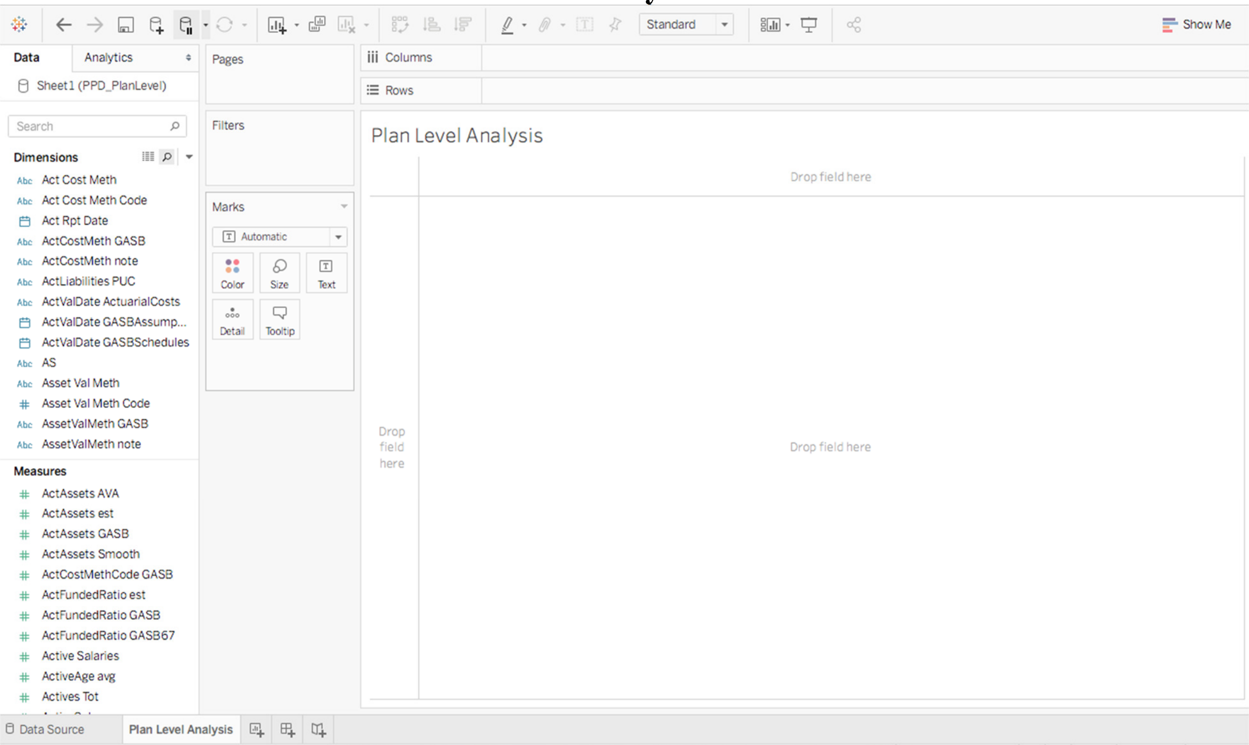
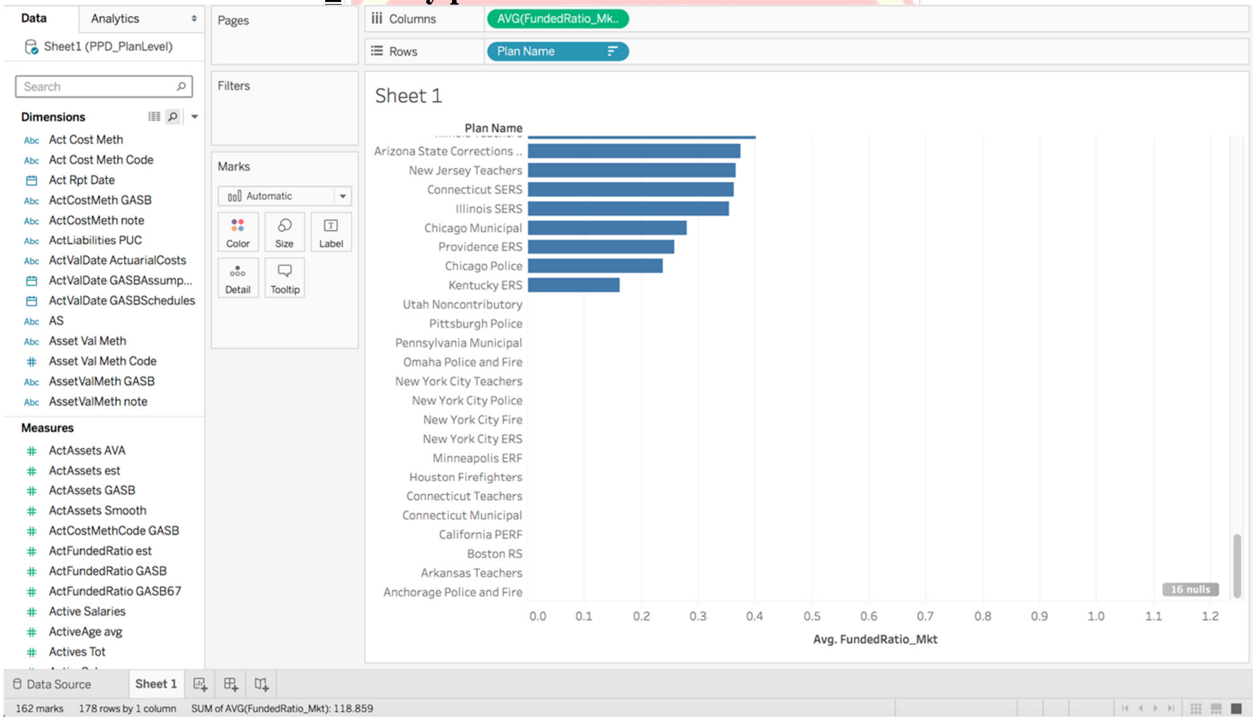
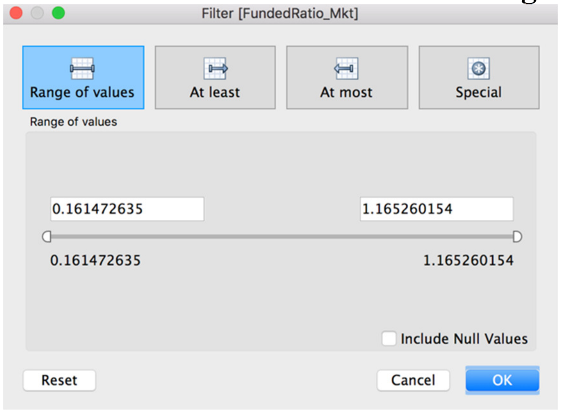


Exhibit 5. FundedRatio_Mkt by plan



162 marks 178 rows by 1 column SUM of AVG(FundedRatio_Mkt): 118.859

Exhibit 6. Tableau interface for editing the FundedRatio_Mkt Filter



Filter [FundedRatio_Mkt]

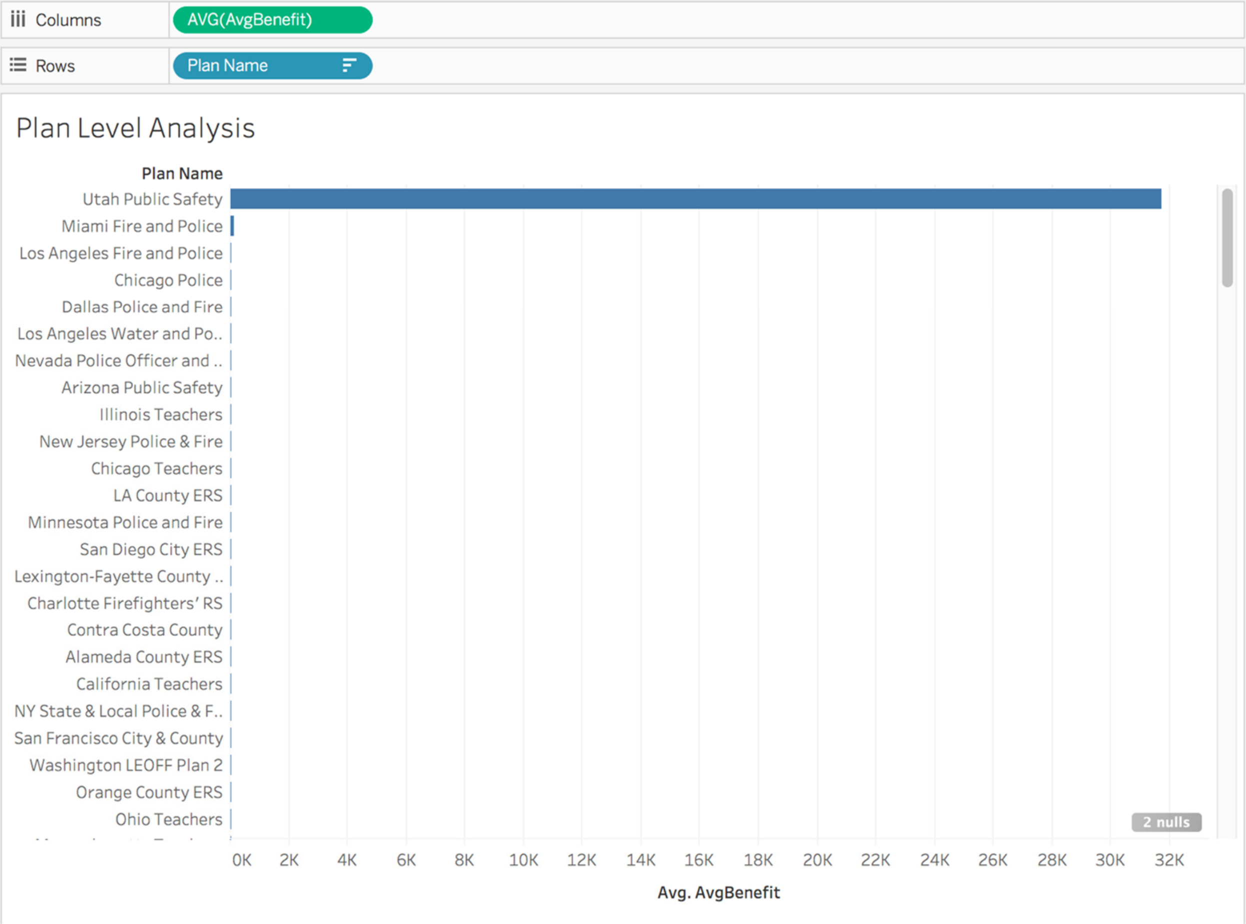
Range of values

0.161472635 1.165260154

0.161472635 1.165260154

□ Include Null Values

Exhibit 7. Average benefit by plan.

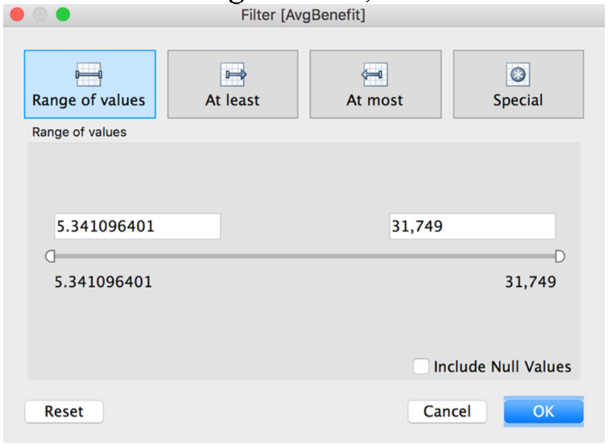


Columns AVG(AvgBenefit)

Rows Plan Name

Plan Level Analysis

Exhibit 8. Average Benefit, Before filter



Filter [AvgBenefit]

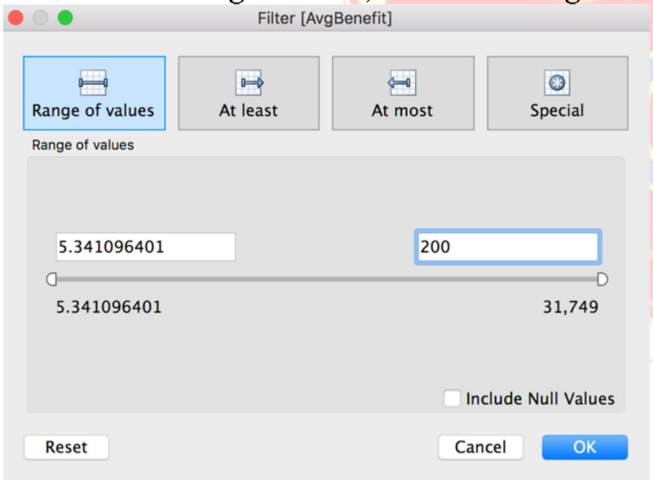
Range of values

5.341096401 31,749

5.341096401 31,749

□ Include Null Values

Exhibit 9. Average Benefit, After Filtering



Filter [AvgBenefit]

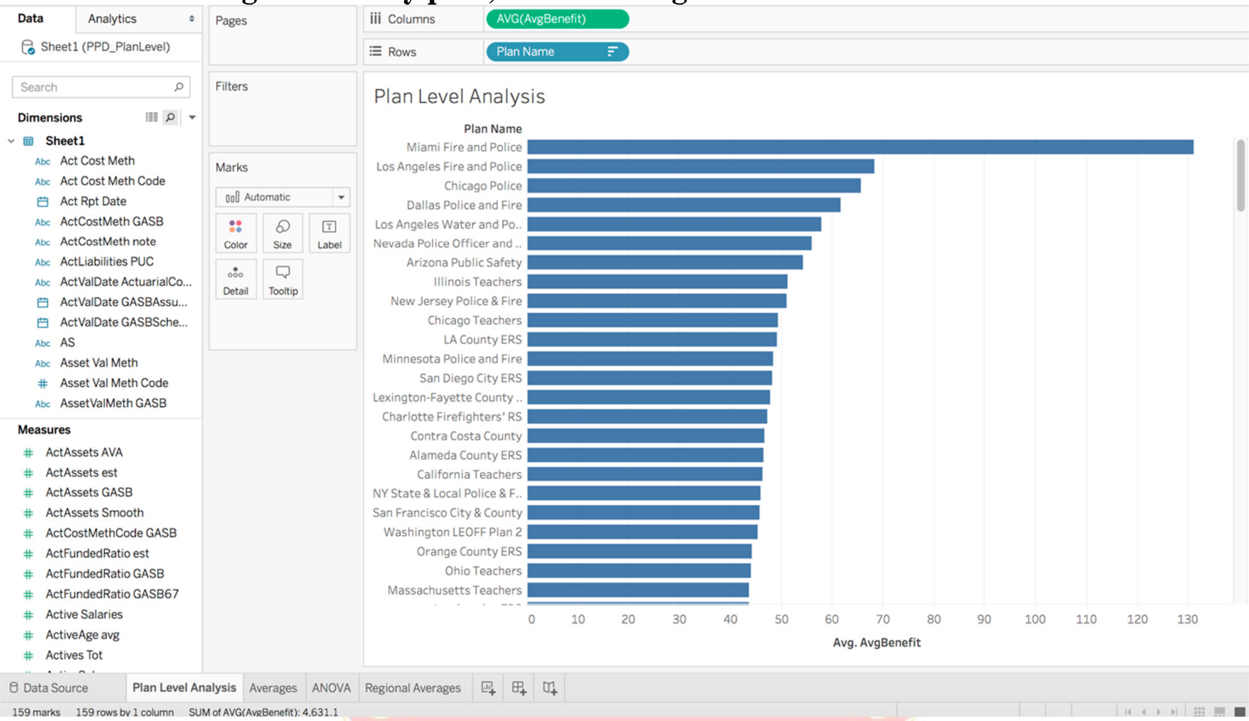
Range of values

5.341096401 200

5.341096401 31,749

□ Include Null Values

Exhibit 10. Average Benefit by plan, after filtering



159 marks 159 rows by 1 column SUM of AVG(AvgBenefit): 4,631.1

Exhibit 11. Label (Text) icon in the Marks panel.

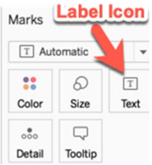
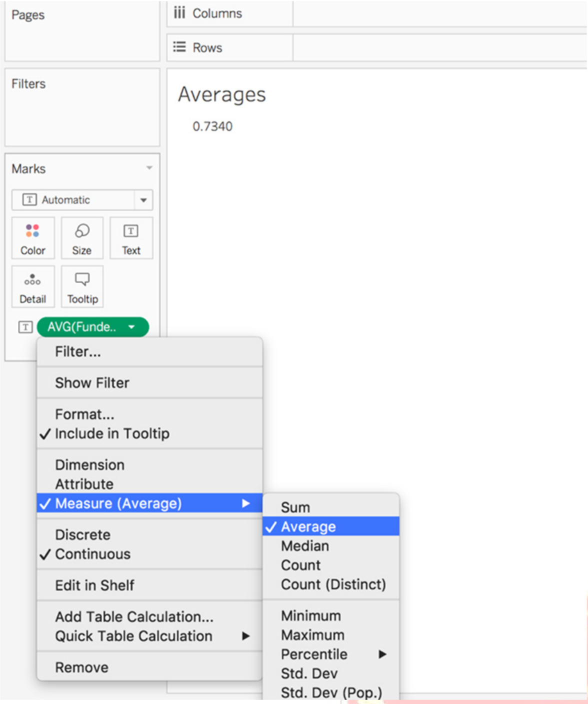


Exhibit 12. Tableau interface for national average of FundedRatio_Mkt



Averages

0.7340

Exhibit 13. Tableau bar chart for state averages of funded status

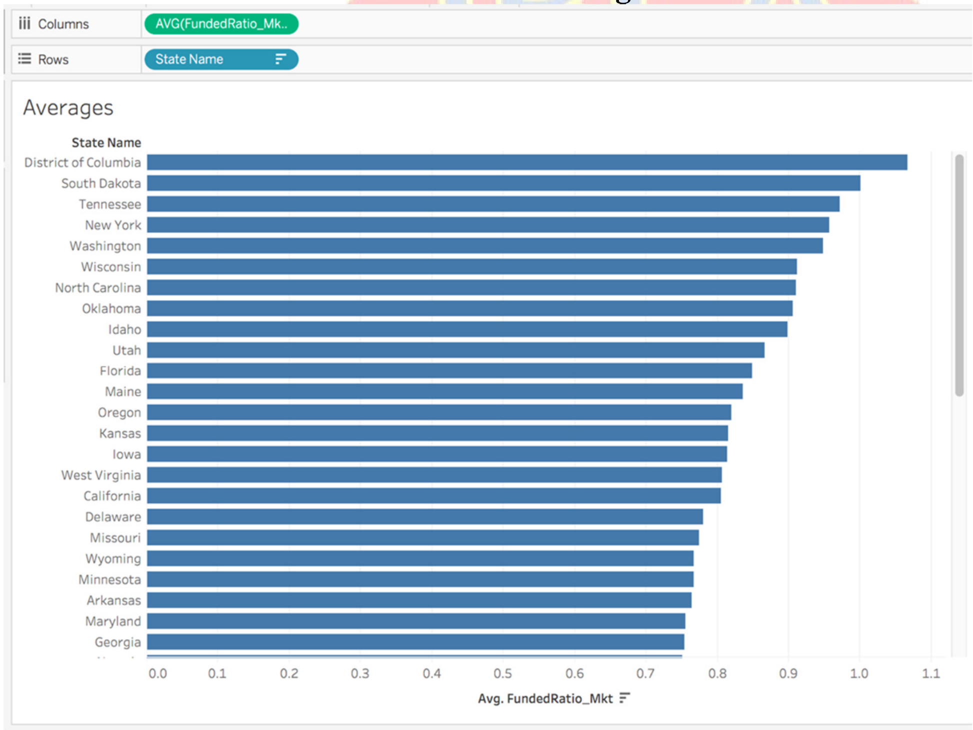
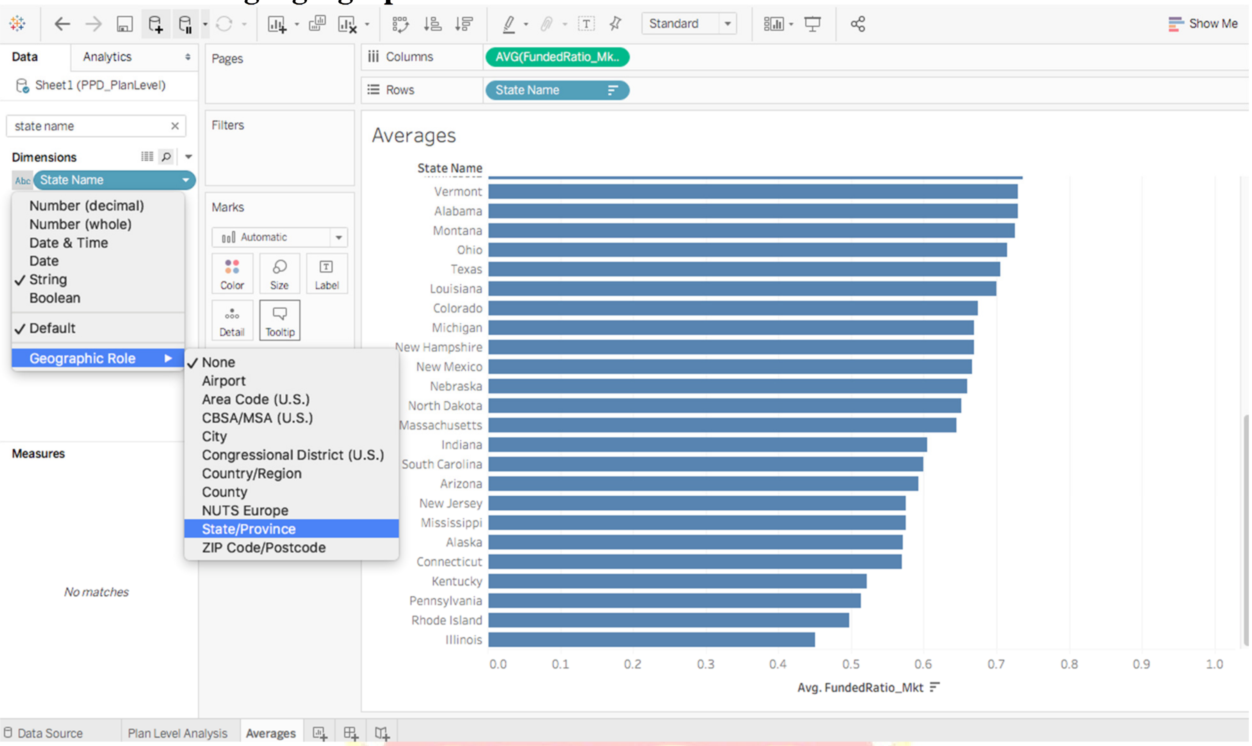


Exhibit 14. Adding a geographic role to state name



Columns: AVG(FundedRatio_Mk..

Rows: State Name

Exhibit 15. Map showing state average of funded status

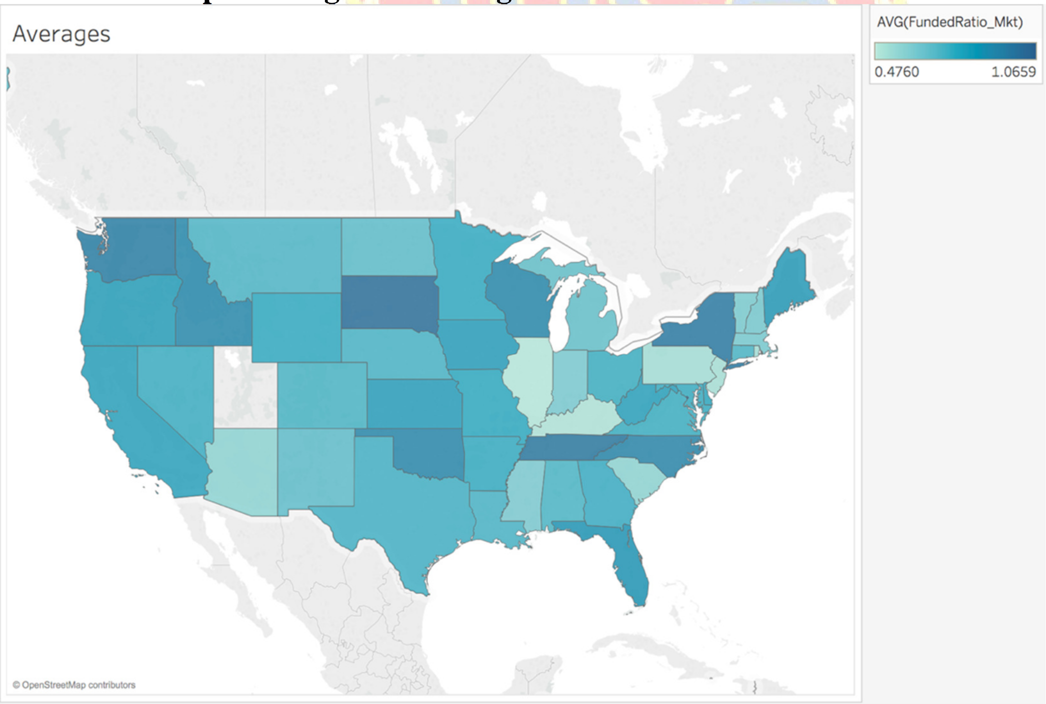
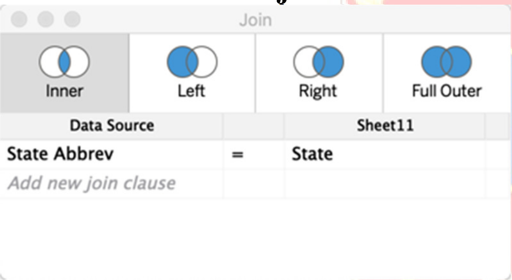


Exhibit 16. List of state abbreviation and its corresponding region for all states in the United States

State	Region
AK	West
AL	South
AR	South
AZ	West
CA	West
CO	West
CT	Northeast
DC	South
DE	South
FL	South
GA	South
HI	West
IA	Midwest
ID	West
IL	Midwest
IN	Midwest
KS	Midwest
KY	South
LA	South
MA	Northeast
MD	South
ME	Northeast
MI	Midwest
MN	Midwest
MO	Midwest
MS	South
MT	West
NC	South
ND	Midwest
NE	Midwest
NH	Northeast
NJ	Northeast
NM	West
NV	West
NY	Northeast
OH	Midwest

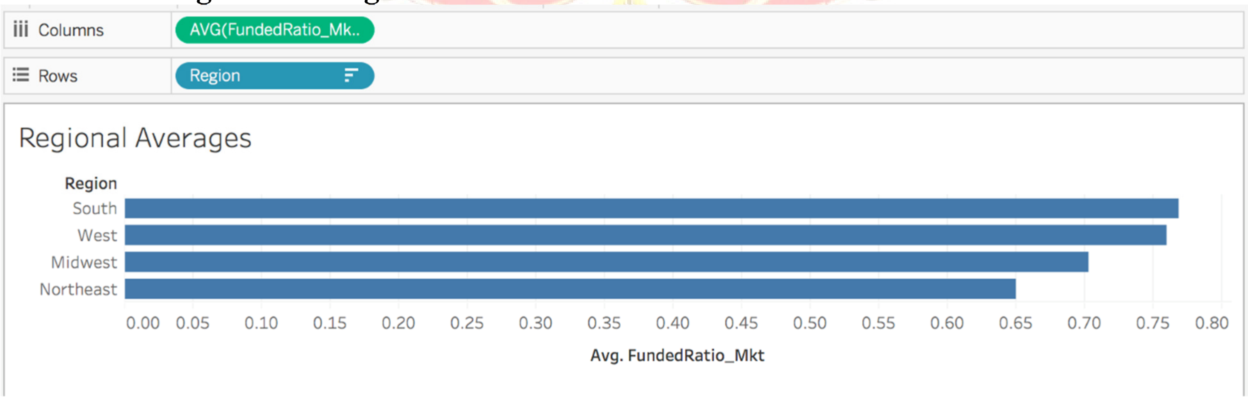
OK	South
OR	West
PA	Northeast
RI	Northeast
SC	South
SD	Midwest
TN	South
TX	South
UT	West
VA	South
VT	Northeast
WA	West
WI	Midwest
WV	South
WY	West

Exhibit 17. Tableau join box



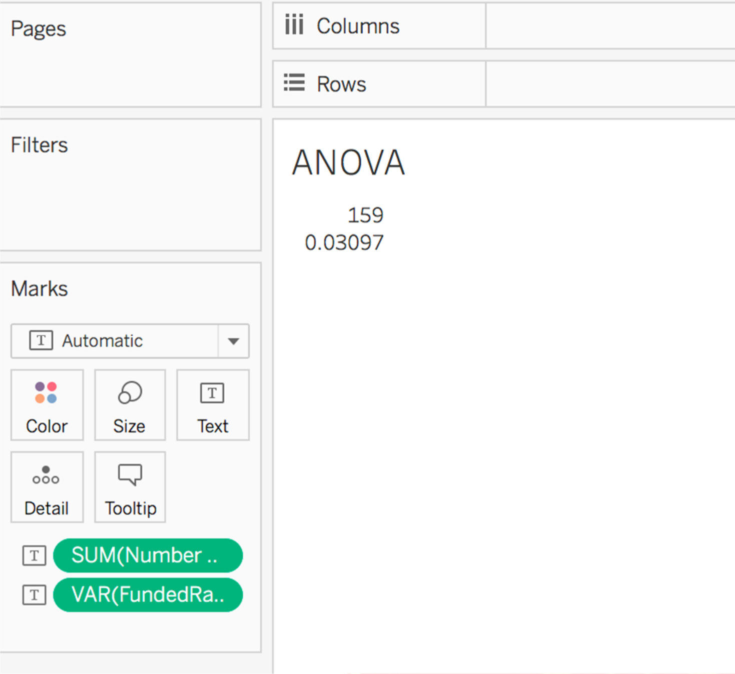
Data Source		Sheet11
State Abbrev	=	State

Exhibit 18. Regional averages of funded status.


Columns: AVG(FundedRatio_Mk..

Rows: Region

Exhibit 19. Total sample size and total variance, in Tableau.



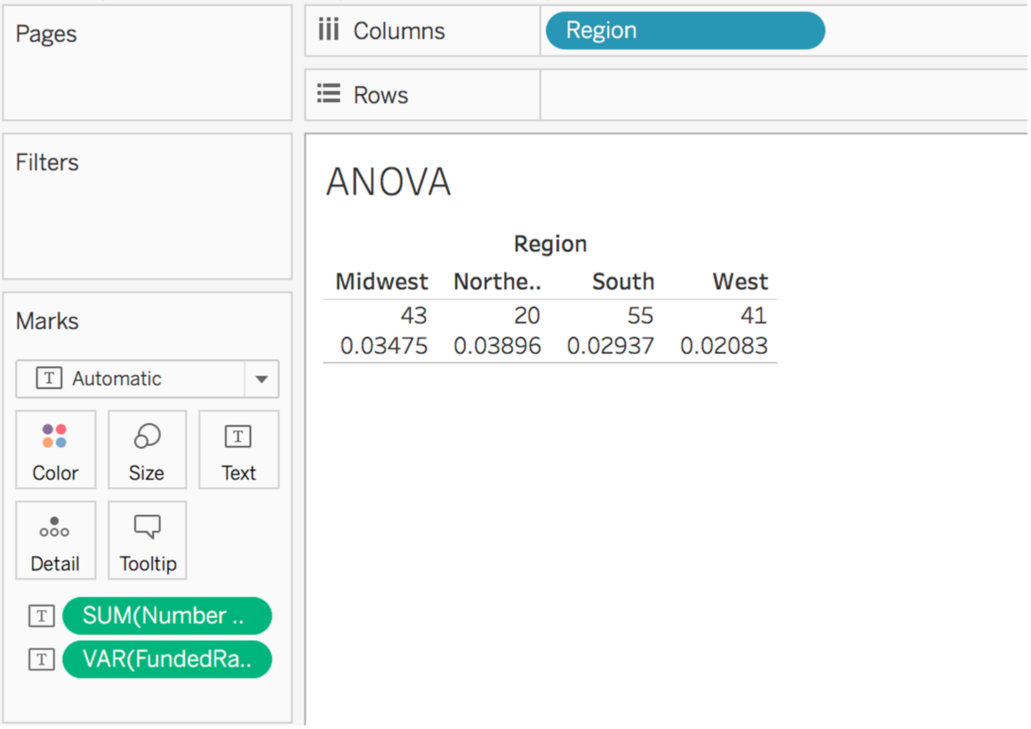
SUM(Number ..

VAR(FundedRa..

ANOVA

```
  159
0.03097
```

Exhibit 20. Group sample size and group variance, in Tableau.



Columns: Region

SUM(Number ..

VAR(FundedRa..

ANOVA

Region			
Midwest	Northe..	South	West
43	20	55	41
0.03475	0.03896	0.02937	0.02083