**Information Visualization: How Data Label and Scale change the speed of comprehension in scatter plot.**

**Name- Prateek Shrivastava**

**1. Proposed Hypothesis (What Question Are You Asking?)**

Explain the question you are asking in your experiment and why it is important?

The primary questionIare asking in this experiment is how the persuasion on the scatter plot changes when additional dimension is added as a data label in the graph.

The result of our experiment highlights how the data label plays a vital role while showing the multi-dimensional data. Also,Idiscuss how the scatter plot can show the data story more effectively when it is presented to non-technical audience.Iare depicting some points here why data label is important:

The data or text labels in the graph give the visual clues to the audience which help the data tell a story blatantly.

It helps in showing some additional information when other important data is already shown in the other axes. Basically, it shows the precise information of the data points.

It enables the graph to stand alone with primary information whereas the supporting information are kept aside.

To use the data label is because it labels the information which comes from the other attributes so if wrong attribute gets selected as a data label then the whole graph will mislead its information.

Also, the length of the data label plays an important part while selecting the attributes in multi dimensional data set because the length of the data label should not overlap with another label. As people always tend to ignore the graph which have too much of text written inside the graph. So, for maintaining the simplicity and tidiness of the graph small text label would be preferable.

The other important feature of the data label is to articulate the important message by the explanatory data label so viewer understand the gist of the graph in the first sight.

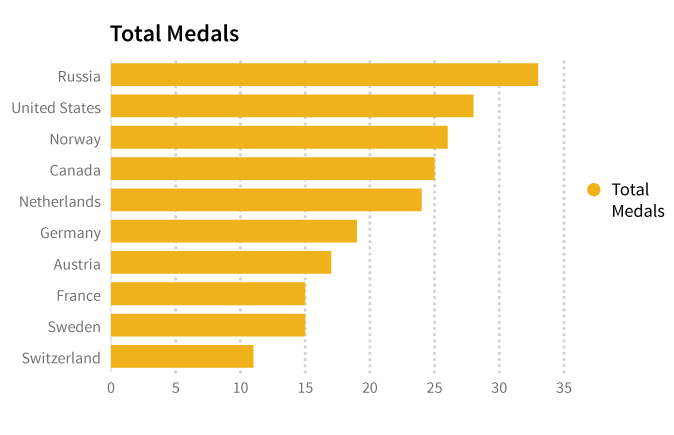
In addition to that, the secondary question of our experiment is how the size of the object (circle) in the scatter plot makes big difference in visualization. As scale is very important in the scatter plot when data are scattered on the basis of size. If 2 different attributes are plotted by the size of some objects like circle or triangle then their relatively size difference should be the primary concern. The data shown in either very small size or very large size makes a bad visualization. As many research have been done in the scale of the plot,Ihave taken one example here which is shown in source 3 below.

Here are some relevant examples and works that show how the data label and object size important in visualization.

1.Source- <https://infoactive.co/data-design/ch14.html>

The examples taken in the above website shows the number of Olympic medals won by country at the Sochi Winter Olympics in 2014. They have done the comparative analysis in between the data labelled graph and non-data labelled graph. Here, the first graph is shown which is not having the data label attributes:

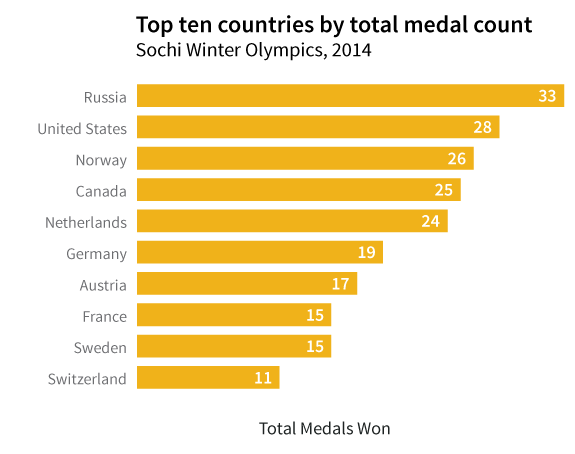
**Without Label:**



So, as per the above graph, it is clear that it is plotted between 2 attributes which are country and Numbers of medals.

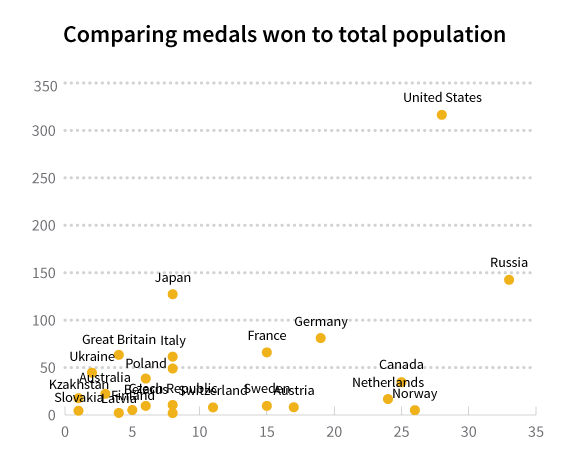
While the other graph with the data label shown below:

**With Data Labels**



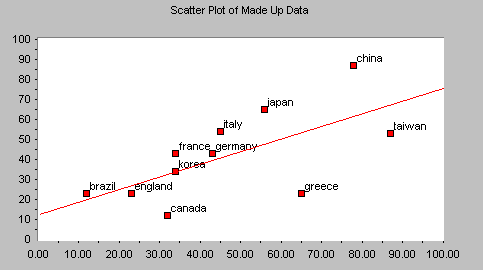
As both graph show the same information and being plotted between the same axes. But it is quite persuasive that the first graph shows the limited information on the given data set while the second graph shows the same data set more didactically. In the second graph, any non-technical users can easily get atleast a clue on the type of story which the data set trying to tell.

Also, one important aspect described in the given example is problem with the text overlapping in data labels. As per the scatter plot shown below the data label are mixed together which leads to confusion and poor visualization.



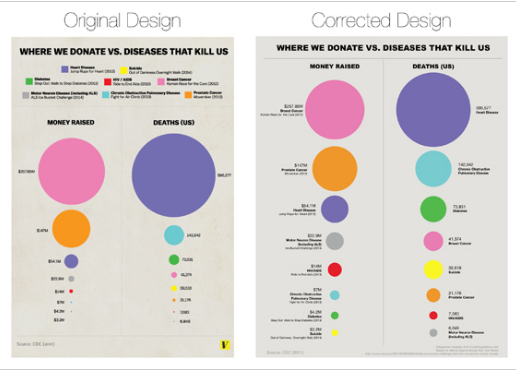
2. Source- <http://www.skymark.com/resources/tools/scatter_plots.asp>

To avoid the overlapping of the text in the data labels, it is important to select the correct attribute. As in the below graph country name are showing clearly because data are scattered with good space.



Source 3. <http://blog.visme.co/bad-infographics/>

As per the comparison illustrated in the below graph, the wrong scale results the smaller size of object in the graph doesn’t reflect the accurate information as it is very hard to understand what dimension is visualized on the basis of size. The graph shows the information accurately as the scale is in the proper order.



**2. Experimental Method**

How will you run your experiment? What are the dependent and independent variables? Are there any confounding variables? How will your design take account of confounding variables? What will the experimental conditions be? What is your overall experimental design (within-group or between-group)? Why?

## 2.1 Overview

In our experimentIdecided to perform the control experiment within the groups soIcould get different range of data which help to get more accurate results. In our study,Iwill recruit 20 participants with different backgrounds. The kind of participant are shown below:

- who all are cricket lover and watch cricket frequently

- who all are not acquainting with the cricket and have never watched cricket.

- who have prior knowledge of graphs.

- who don’t have any knowledge of graphs.

We will mix them in a group on the random basis and will ask them to come into the room where they be see the graphs and answer the questioner displayed in the computer screen. Then,Iwill collect their answers on the set of questionIhave asked during displaying of the graph. Now, to be able to conduct the test successfullyIdecided our measurable variables as follows:

**Dependent Variable** - The accuracy of answered question in this experiment.

**Independent Variable** - Data label and scale of the graph to visualize the data.

**Confounding Variable** -Ialso considered the confounding variable to increase the accuracy of the result as confounding variables are outside influence which changes the effect of dependent and independent variable. They can also adversely affect the relation between the independent variable and dependent variable which may result the false correlation between the dependent and independent variables. In our experiment confounding variable may be:

* Eye sight of the subject.
* Ability to understand the language in which the text written as a data label.
* Understanding on the size of the objects (circle, bubble) in the graph.
* Familiarity with the scatter plot shown in the computer screen.
* Positioning of the screen placed into the room as if screen too far from the viewer then size of the object may be inappropriate. Since it our resultant measure so it should be precise and should not be misconstrued by the external effects.

So,Ihave to design the experiment in a way such that the confounding variables were accounted for. This is an achievable task sinceIare conducting a controlled experiment and are in position to choose our audience, the venue for the experiment and can conduct any prior session on some information related to the prerequisite if needed.

We will take the following measures to account for the confounding variables:

Firstly,Ibegin with the pilot study where participants are asked to answer the set of question which is related to English language, graph knowledge, understanding of the computer screen display. Also, participants have to go through some eye sight test where they have to identify different font sized characters display from the same distance which they will see in the actual experiment. So, after completing the pilot study,Iwill observe the result and decide who are the participant eligible to be a part of the final group on which actual experiment will perform. So, by this study, people with low eye sight and poor knowledge of general English will be eliminated. However, the remaining participant which are good in English language and have good eye sight but have considerably low knowledge of graphs and computer screen display will be gone through one training session.

Besides,Iwill give all participants the same resolution computer screen and same positing of the monitor so they will have the same experience.

So, this is howIwill avoid the intervention of the confounding variable in our experiment.

## Data collection

In our experimentIare focus on how the different graphs change the persuasion of the same information. So,Ineed to find the accuracy of the different graphs of the same data. In order to calculate the accuracy,Icollect feedback from the audience on the basis of the questionerIhave given in our experiment and analyse which graph shows the best representation of the data. The answers provided by the participants are based on their experience with the kind of visualization presented to them. As some of the participant have seen the graph with the label, some of them have seen the graph without label and other have seen the graph with different scale of the scatter plot of the same data. So,Ican analyse which graph are being best in representing the information on the basis of number of correct answer. Also, it is equally important that which category’s participant give more correct answer.

Data collected during the experiment of participant’s performance and the question on which have judged are given below:

* We will gather the information about the speed of noting the inquiries said in the representations in area 3: Data Visualizations, and in addition the responses to every one of the inquiries, expressed.
* There will be a blend of subjective and objective measurement in our experiment as the objective measurement depends on how well individuals perform in the task, independent of what they encounter while performing in the task so the quantity of right answers in our experiment is an objective measurement.
* Whereas the Subjective estimation, imply to measures that have to do with actually experience participants have got during the experiment. SinceIhave questions that rely upon the participant's personal experience, for example, simplicity of perusing the perception and speed of replying,Iwill utilize a blend of the both two measures to evaluate the performance.
* The resultant data of this experiment will describe how the same scatter plot with the different visualization can change the people’s way to understand the information plotted.
* The multi variable graph shows considerably different information if the size of the scale changes. Also, how data label plays important role in showing the information.
* As our proposed hypothesis cites, how the persuasion on the scatter plot changes when additional dimension is added as a data label in the graph. And how the size of the object (circle) in the scatter plot makes big difference in visualization. This experiment helps in supporting our hypothesis by checking the quantity of correct answers and speed of replying to the questions. Also, observing the way people answering the question to know how many have answered just superficially.

## Selected subjects

In our experimentIhave decided to take the participants from largely diverse backgrounds. In the subjectIhave select consists 4 type of people which are cricket lover who watch cricket frequently, non-cricket lover who have never watched cricket, Statistics student who knows graphs and other people who don’t have any knowledge of graphs. Since UCD has all kind of student and staff member which is needed in our experiment soIhave decided to select the people from the UCD itself. However, in the pilot study with an extra member,Iwill guarantee that the overall procedure runs easily and that the tasks are straightforward. Before the actual experiment,Iwill also consider the eye sight test.

The representational sample is gathering of members chose from a bigger populace that nearly coordinates the qualities of the populace overall.Ican state that the specimen is a genuinely exact reflection of the populace from which the example is drawn. So, the group of 20 people selected after the pilot study will be the representative samples. SoIwill have all necessary kind of people in our experiment andIcan better generalize the result to the population as a whole.

The members will be sourced from University College Dublin. AsIare searching for subjects from variety of backgrounds,Iwill search for volunteers among the student body such as student unions, workforce staff, and individuals who work in a non-academic job.Iwill promote our experiment by sending messages to students and staff and by leaving notices in various territories of grounds. Out of the volunteers,Iwill then select 20 individuals who compare to our criteria and speak to an extensive variety of members.

## Data analysis

We will be analysing the information gathered amid the experiment in the following ways.

The accuracy of the answer given by the participants on the questioner presented during the experiment. The speed of the answer provided while looking at the graphs. The observation on how the participants answer the question asIhave to observe whether they have answered from the seeing the visualization or just made flukes decisions.Irecord the effect of each graph on the answers.

After completing the experiment,Iwill group the 3 graphs used into the experiment. And then analyse them in the following manner:

The first graph which is the scatter plot of the cricket data set having 4th variable as a data labels and properly scaled the size for the 3rd variable.Iwill count how many participants have given the correct answer while looking at this graph.

The second graph is a scatter plot with no data labels and have improper size of the object.Iwill count the answer for this graph.

The third graph is a scatter plot with no data labels and no sized variable. Will be take the answer of this graph as well.

After analysing the overall result, the graph on which the highest number of correct answer came will be the best graph among all.

Also, by checking the individual results,Ican conclude that which graph is the best for which group of members.

This analysis helps answer the questionIinitially proposed on the grounds that:

It bolsters us to build up the connection between data labels and scale and their effectiveness in the information by giving us how utilizing the data labels and scale affect the appropriate responses and also the speed of giving the right reaction asIrecord the time in which the questions are being addressed and addressed correctly.

The accuracy of the answers and speed of the response describe which is the best visualization graph which were our primary agenda.

## Practical setup

To perform our experiment,Iwould follow the below steps:

1. **Pilot Study**

BeforeIrun the final experiment,Iwill perform the pilot study. The pilot study is used to ensure that the measure used in the final experiment are gathered correctly. AsIhave confounding variable too soIhave run the Pilot study first to control these variables. In the pilot study participants will be asked to answer the set of question which is related to English language proficiency, graph knowledge, understanding of the computer screen display. So, the result of the pilot study will give the appropriate participants for the actual experiments.

Besides this,Iwill conduct one eye sight test for checking the quality of eye sight of the participants. In this test, participants will be given the range of different font sized characters and numbers from the same distance which they will experience in the actual experiment. This will give us the number of participant with the required eye sight.

In addition to that,Iwill conduct a training session for the participant who have considerably low knowledge of graphs and computer screen display.

1. **The actual Experiment**

In our main experimentIwill ask the participant of 20 people group which are selected after the pilot study to answer the question displayed in the screen. Since the experiment will be conducted online, each participant will be allocated one computer in the lab. The resolution and positioning of the screen will be same for all the participants so each participant will get the same experience. Every participant will see one graph which will be choose on the random basis and answer the question shown in the questioner next to the graph.

As the experiment is in within group and participant will get different graphs so outside influence will be minimal.Iwill lead our experiment in a well and consistently good length room in which each subject will be seen in all treatment conditions.

The below are the instruction given to the participants:

* The members will be made a request to complete a task while being watched and their reactions will be recorded and measured.
* They will be told to not to talk to other members amid the test. And ask to keep their choices private and not to uncover their decisions to some other member amid or after the analysis.
* They will be asked to be situated in front from a Computer and will be informed that the whole test is to be finished in fifteen minutes.

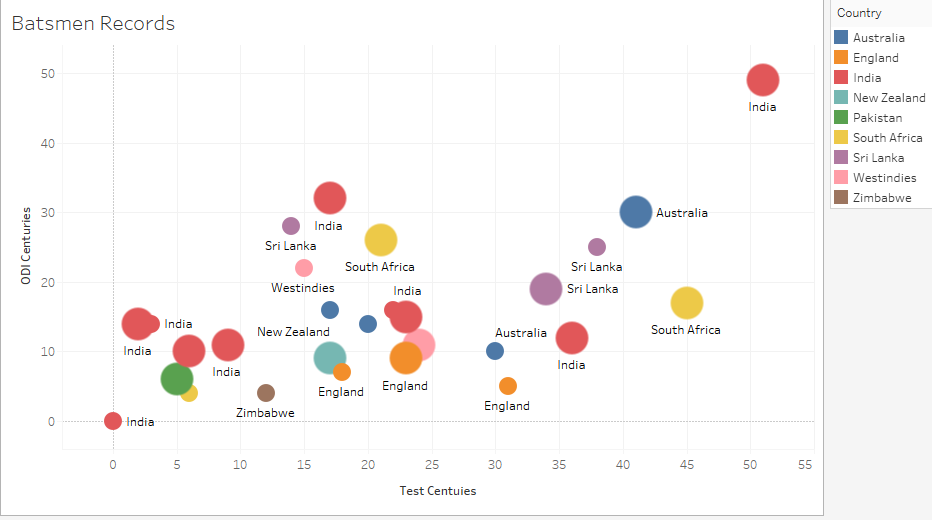
We will reassure the participants that our goal is to recognize conceivable faults in the idea of use of data label in the scatter plot and the importance of the scale in data visualizations, and not to test the participant’s own particular capacity or insight and that it isn't a reflection all alone capacity.

# 3. Data Visualisations

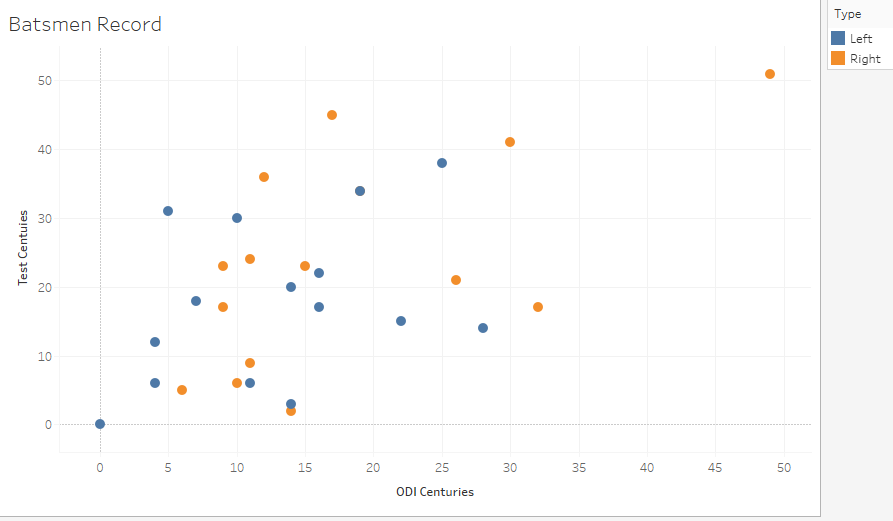
The graph whichIare be using in our experiment are given below:

In the first visualization,Ihave plotted the 4 variables, on the x axis Test Centuries, on the y axis ODI Centuries, size of the object is showing the type of the batsmen (right handed or left handed) and the data label are showing the country name of the players.

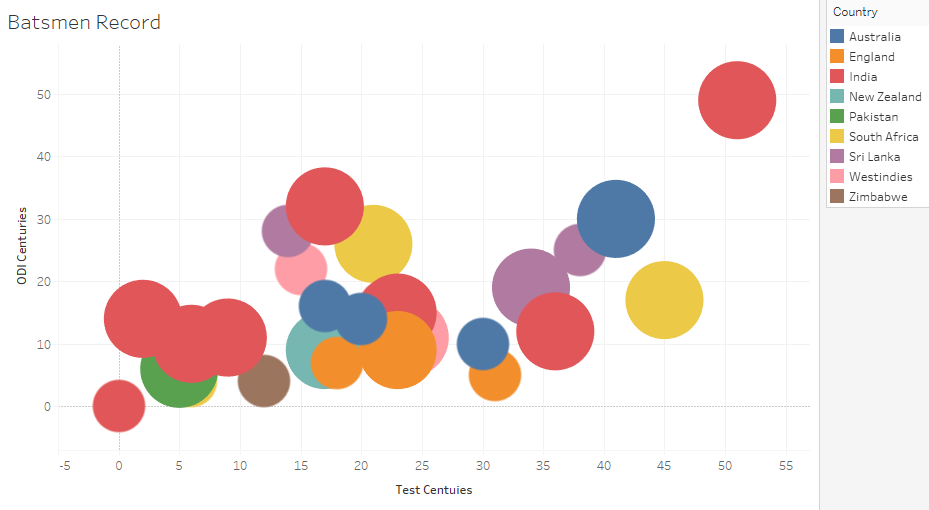
In this graph all the 4 variable are shown in a very proper manner and showing the information blatantly.



The second graph plotted on the same cricket data set having the 4 variables but the county name is shown in the description attribute. So, when someone brings the mouse arrow on the circle, the county name will be displayed in the description.



The 3rd graph which also have the 4 variables of the same cricket data set but have the poor scale of the circles. Also, the 4th variables i.e. county name is shown in the description attribute. So, when someone brings the mouse arrow on the circle, the county name will be displayed in the description.



The questionerIhave presented in our experiment is given below:

Q1. Which type of the batsmen have scored the highest centuries in ODI and Test cricket?

1. Right
2. left

Q2. Which type of the batsmen have scored the test century more than double of the ODI centuries.

1. Right
2. left

Q3. Which type of the batsmen have scored equal number of centuries in ODI and Test Cricket.

1. Right
2. Left

Q4 which country have produced large number of left handed batsmen.

1. India
2. Pakistan
3. South Africa
4. Australia

Q5 Which country have more batsmen with large number of ODI centuries scored batsmen.

1. India
2. Australia
3. South Africa
4. West Indies

Q6 Which country have more batsmen with large number of Test centuries scored batsmen.

1. India
2. Australia
3. England
4. West Indies

Q7 Which county has larger number of centuries scored batsmen in both of the format.

1. India
2. Australia
3. England
4. West Indies

Q8 Which county has highest number of centuries scored batsmen in both of the format.

1. India
2. Australia
3. Pakistan
4. West Indies

Q9 Which county has lowest number of centuries scored batsmen in both of the format.

1. India
2. Australia
3. America
4. West Indies

Q10 Which county and type have the number of ODI centuries just one more than the Test centuries

1. South Africa
2. Australia
3. England
4. West Indies
5. Left
6. right

These are the questionerIhave used in our experiment to analyse how the use of data label and scale of scatter plot change the persuasion of the information.Ihave analysed the data in 2.4 on the basis of these question which are answered by the participants.

# References

I have created my own data from the various cricket related website which are given below:

<http://www.espncricinfo.com/icc-cricket-world-cup-2015/content/squad/817409.html>

<https://gigadom.wordpress.com/2015/02/06/analyzing-crickets-batting-legends-through-the-mirage-with-r/>

<https://link.springer.com/chapter/10.1007/978-1-4020-6528-6_6>

<http://www.lithoguru.com/scientist/litho_papers/JM3%20editorial%202014%20q1_Figures%20part2.pdf>

<https://infoactive.co/data-design/ch14.html>

<http://www.skymark.com/resources/tools/scatter_plots.asp>

<https://en.wikipedia.org/wiki/Cricket>

The data set I have used in plotting the graph is attached below:

