Visualization Lies - Influencing Visual Perceptions using Biased Visualizations

Prepared By

Shyam Mohan Kizhakekara shyam.mk@outlook.com

1 Proposed Hypothesis

The proposed work revolves around the question "Can certain skewed or unconventional visualization techniques actually mislead the audience, in terms of drawing inaccurate or incorrect conclusions and shift their focus to the biased area in the graph?". Through our experiment, we try to demonstrate how biased data visualizations can have impacts on the perceptions of the target audience. Visualization Lies is a terminology that describes how graphical misrepresentation of the data done either intentionally or unintentionally can influence viewer perceptions. Several such methods exist, and they are resorted to, when there is a need to emphasize on a specific component or section of the graph.

The major importance of this experiment lies in understanding the extent to which these biases are useful in impression management. Impression management is used to influence people by regulating or controlling the information being presented to them. This being a wide topic and applicable to various fields, we try to narrow our focus to a few of them.

- **Business & Corporate Brand Building** In business, the strategy of portraying transparency by presenting information of maximum value to the concerned stakeholders can be useful in gaining their trust, thereby providing provisions to expand the business and eventually enhance the brand value.
- Politics & Development Indices Political impression management is a necessity today, as social media has an extremely high influence on election campaigns and results. Furthermore, the growth and development indices of countries are considered as baselines to evaluate the performance of ruling governments today. Hence, distorted graphical representations can be of help here as well.

Statistical lies are also used in false advertising, profit-loss margins in sales, stock market variations etc., in order to manipulate the intended audience.



Figure 1: US Smartphone Market Share (2008)

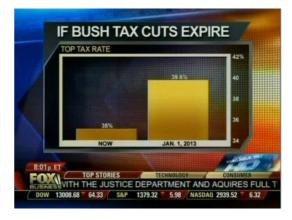


Figure 2: Fox Chart - Bush Tax Cut Edition (2012)

For instance, during the keynote address at WWDC in June 2008, Steve Jobs presented a pie-chart representing the US Smartphone Market-share. Apparently, the Apple shares depicted in the chart got the closer perspective, and somehow the section representing 19.5% ended up with a larger surface area than the section representing 21.2% of Others' shares, as shown in *Figure 1*. In this case, the chart has been tilted backward thereby increasing the surface area for the Apple segment as compared to that of the segment for Other. Another notable observation is that the market shares are illustrated by the brand and not the OS itself. Through these defied conventions, Jobs managed to bring the focus of the audience to the share of Apple, even though it has a smaller share compared to other brands [1].

Similarly, in 2012, Fox News portrayed the change in the top tax rate if the Bush tax cuts expire, in such a way that the rate in 2013 would be around five times higher than the rates now, which was termed dishonest by many (Refer *Figure 2*). A mere difference of 4.6 points was shown as a huge difference by truncating the y-axis of the graph i.e., the staring value of y-axis was pointed to a value just below the smallest bar, if the values would be represented in a bar graph. This resulted in a misleading representation that the tax cut rates would be very high in the upcoming year ^[2].

In the corporate world, building a brand by influencing the impressions of the stakeholders, about the organization is common. This is done through the extensive usage of biased techniques in corporate storytelling. Similarly, the amount of distortions in graphs depicting financial data are also quite high. Some of the existing research concludes that audience with poor levels of knowledge in financial domain are more likely to be misled [3][4][5] or even the ones with a good level of understanding can be manipulated [6]. This conclusion can be applied in general as well, as the non-specialist observers in any field are easier to be influenced than the specialist ones.

2 Experimental Method

2.1 Overview

The proposed work is to understand and rank the misleading visualization techniques based on the extent to which the target audience is influenced by them and how those visualizations manipulate their inferences. There are several different techniques available for creating biased visualizations and we make use of the following ones in our experiment.

- Cumulative Measures
- Tilted 3D Pie-Charts
- Omission of Data Points
- Inverted Axes
- Truncated Axes
- Map Based Distributions
- Comparison based on Different Scales
- Irregular Axis-Scales
- Stacked Area Graphs
- Bubble Charts

The idea is to present a mix of both accurate and misleading visual representations of the same scenarios to a selected set of people through an online survey and record their interpretations. The recorded results of the survey would be used to deduce the final conclusion of the experiment. We intend to select 100 individuals with different backgrounds and expertise, as the subjects for this experiment.

In our experiment, the independent variables would be the different types of graphical representations using the above-mentioned techniques. The dependent variables would be the accuracy or correctness of the interpretations made from the graphs, by the chosen subjects. The confounding variable in the scope of this experiment would be the ability of the participants to perceive the graphical representations and derive inferences from them, along with their knowledge or expertise in the respective field.

The online survey would be developed using Microsoft Office Forms and the URL for the survey would be shared with the participants through email. The answers of the participants would be stored on the server which would later be analysed to reach the desired conclusion. The experimental conditions defined for our proposed work are as follows:

- Choose the subject set by using the random sampling technique.
- In order to nullify the effect of the confounding variables identified, we make sure that:
 - None of the selected participants are colour blind.
 - ➤ All participants possess the basic ability to interpret a graph and infer some observations from it.
- All participants must know how to use a web browser and access a URL.
- All participants must have prior experience in being a part of online surveys.
- All participants must possess or have access to a computer or a smartphone, using which they can take the survey.
- The participants are required to take the survey in a calm space devoid of any distractions.
- A minimum of 100 responses must be obtained through the survey.

The whole experiment follows a 'within-group' design, as it aims at analysing the difference in perceptions within a group of people, based on the visualizations presented to them. At the same time, the selected group of participants would have enough variety in terms of the knowledge about the domains the graphs represent, and the ability to interpret the visualizations.

2.2 Data Collection

The survey would comprise of 11 different visualization scenarios where each scenario would have an accurate representation as well as a misleading representation. The answers given by the participants for the questions based on the graphs in the survey would be stored on the server, as one entry for each participant which would be identified by a unique identifier.

We plan to collect a mix of both objective and subjective measurements from the result set. The objective measure here would be the evaluation of how fast the survey was completed by the participants. This would be deferred from the time taken by each participant to complete the survey and submit the results. The inferences drawn by the

participants from the graphs would be our subjective measure i.e., the options selected by a participant as the answers to the questions in the survey.

In order to prove our hypothesis that people can actually be misled by introducing certain skewed or biased visualization techniques, we need to collect these data. Our objective measure would help us short-list the 100 responses which were filled in the shortest time. We short-list the 100 fastest responses based on the assumption that faster the response rate, higher is the probability of capturing the initial perception of the participant. The subjective measure would give us the information about the perceptions of each participant regarding the graphs shown to them and the comparison of their answers for the accurate and misleading instances would help us understand if their interpretations were actually manipulated. Next, based on the number of participants who were misled, we would figure out which techniques influence the perceptions of people the most.

2.3 Selected Subjects

For our experiment, we would be randomly selecting 100 individuals from different age groups and social backgrounds, with varied expertise and interests and are not colour blind. Some of them would be specialists in the areas of business, economics, finance, politics, civics and international relations while some of them would possess a fair understanding of these domains. The rest of them would be from other streams who would not have much knowledge about these domains. The aim is to have a good mix of people, which consists of domain experts, people with a good understanding of the domain and non-specialist audience.

The number of participants in the selected groups would provide enough variety in terms of the opinions and perspectives required for the experiment to be carried out successfully. We have also made sure that the defined size of the subject set is appropriate to make sure that the experiment is completed within the stipulated time and would not cause any issues related to the administration of the project.

Most of the participants would be students pursuing different undergraduate or graduate programmes pursuing different courses, at University College Dublin. Some of them would be selected from the courses related to the domains mentioned earlier and some would be from completely different courses. We would also be involving a few individuals from different parts of the globe with expertise in any one or more of those domains, in the experiment. All participants must meet the experimental conditions defined.

2.4 Data Analysis

The first step is to retrieve the result set from the server and export it as an excel worksheet. Then, we sort the result set in the ascending order based on the time taken to complete the survey and filter out the top 100 responses i.e., the responses filled in the shortest time. As mentioned earlier, the survey would comprise of 11 different visualization scenarios. With two graphs for each scenario, one actual and the other misleading, we intend to analyse the magnitude to which viewer perceptions are manipulated, for each bias technique.

Next, we apply an evaluation technique on the filtered response set by generating the corresponding confusion matrices for each scenario. These matrices would help us in

performing an appropriate comparison between the obtained results of the experiment and the actual ground reality. The confusion matrix would provide us the counts for True Positives (TP), False Positives (FP), True Negatives (TN) and False Negative (FN) categories respectively, that would help us in calculating the number of individuals whose interpretations actually got manipulated in each scenario and the number of individuals who were able to interpret the graphs accurately.

Using the confusion matrices, we calculate the accuracy (fraction of perceptions that are correct) for each technique using the below formula.

$$\label{eq:accuracy} \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Once we have the accuracy data, we proceed to calculate the True Positive Rate and False Positive Rate for each scenario and plot the Receiver Operating Characteristic (ROC) curve between them.

$$TPRate = \frac{TP}{TP + FN}$$

$$FPRate = \frac{FP}{FP + TN}$$

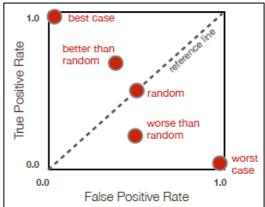


Figure 3: ROC Analysis Skeleton

From the plotted curves, the curves below the random reference line would be for those bias techniques that actually influenced the visual perceptions of the majority of the participants. The most misleading technique would be the curve closest to the bottom-right corner of the plot shown in *Figure 3*. Additionally, this technique would have the lowest accuracy value as well. At the same time, the curve closest to the top-left corner of the ROC graph in *Figure 3* would be the least misleading technique and the majority of the participants were able to draw the correct inference in spite of the graph being biased. We finally create a table that ranks each technique based on its accuracy measure.

From the accuracy table, we would be able to deduce the skewed or biased visualization technique which is the most misleading and shifts the focus of the chosen subjects.

2.5 Practical Setup

The experiment would be conducted online and hence, the chosen subjects would be able to participate in the experiment from their own physical locations, using their own personal computers or smartphones. All participants would be provided with the URL of an online survey through email. The participants are supposed to take the survey in a private room without any visual or auditory distractions. Each of them would have to

respond to the survey and their responses would be recorded with a unique access code, for further analysis and conclusion of the experiment.

Upon starting the survey, the participants would be asked to register by entering some personal details along with their expertise in any of the listed domains. In the next page, the participant would be provided with all the necessary instructions to fill out the survey, clearly describing the total no. of questions, no. of mandatory questions, how to answer them and how to save their answers. Once the answers are saved, the user would be shown a confirmation of the same.

The survey would consist of a total of 22 questions divided into two sections A & B, having 11 questions each. All the questions would be objective in nature i.e., multiple choice questions. Each question will have an associated graphical representation displayed along with it and the participant would be asked a question based on the provided graph. The participant would have to select any one of the listed options as the answer. The 22 questions in the survey would actually be based on 11 different data visualization scenarios, where each scenario would have an accurate representation as well as a misleading representation.

Section A of the survey would have the set of data visualizations that are created using biased visualization techniques, while section B would consist of the accurate representations for the scenarios. There would be a minute-long break between the sections. This is to create a small distraction for the participants so that they can start the second section afresh and would not be able to relate between any two graphs for the same scenario. Also, once the participant enters section B, he/she cannot navigate back to the previous section.

The survey would be created using HTML and Microsoft Office Forms and the results of the survey will be stored on the server, which would later be exported as an excel worksheet. From the results, we would shortlist those 100 responses which were filled in the shortest time. This is done based on the assumption that the participant who has completed the survey in the shortest time, has done it using the first perception he/she had, regarding a graphical representation. Finally, we calculate the necessary parameters and figure out the bias visualization technique that manipulates the perceptions of maximum no. of participants.

3 Data Visualisations

All the relevant visualizations that we have used for this experiment are given below. For each scenario, we have displayed the accurate representation as well as the misleading representation. We have also included an explanation as to why these representations could mislead the chosen set of people and how they would possibly interpret the graphs.

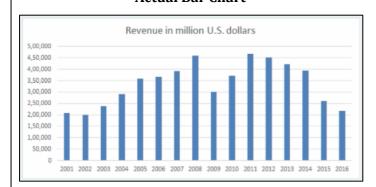
We have used a total of 22 graphical representations, for 11 different scenarios i.e., 11 accurate visualizations and 11 biased visualizations. These are the same graphs that would be presented to the participants in the survey sections A & B. These visualizations were developed using Microsoft Excel and Tableau.

The questions in a sample survey are given in the Appendix (Section 5.2).

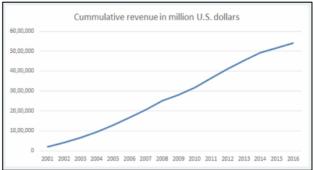
1. Exxonmobils Revenue from 2001-2016

Bias Technique Used: Cumulative Measure

Actual Bar Chart



Misleading Line Chart

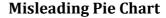


Explanation: In the actual scenario, the overall trend of the revenue of Exxonmobils is decreasing. But the representation of the same scenario using cumulative revenue measures, shows an ever increasing graph which can mislead the audience to a considerable extent.

2. Most Popular Social Media in 2017, by No. of Active Users

Bias Technique Used: Tilted 3D Pie Chart



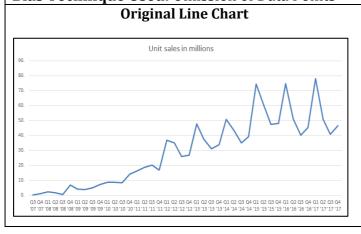




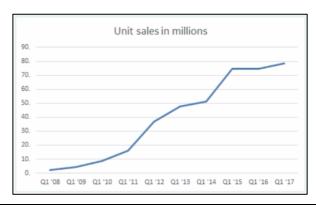
Explanation: When the participants are asked to compare the user traffic on Youtube with the user traffic on Facebook Messenger, there is a high probability that they would perceive the the user base of Facebook Messenger to be equal or larger than that of Youtube. But in reality, Youtube has the higher no. of users.

3. Apple iPhone Unit-Sales Worldwide from 2007 to 2017 (Quarter-wise Distribution)

Bias Technique Used: Omission of Data Points



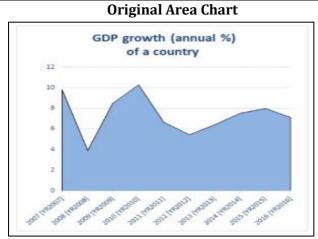
Misleading Line Chart

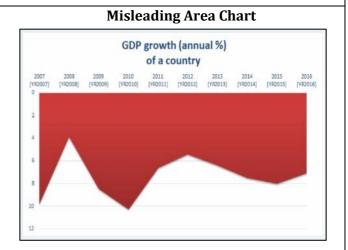


Explanation: The unit-sales for iPhones were increasing year on year but the sales were fluctuating in every quarter, as shown in the actual scenario. The misleading graph shows the revenue growth in Q1 of every year which could lead to a perception that the sales never went down. Usually iPhone sales shoot up in the quarter after the new iPhone launch.

4. GDP growth (annual %) of India from 2007 to 2016

Bias Technique Used: Inverted Y-axis

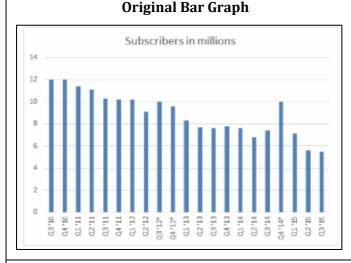


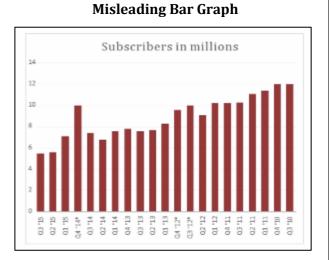


Explanation: The accurate representation of the graph shows a dip in GDP from year 2015 to 2016, but the inverted graph might give a perception to the user that the GDP is on the rise after 2015 i.e., the user presumes it to be an increasing curve.

5. Number of World of Warcraft subscribers from year 2010 to 2015 by quarters

Bias Technique Used: Inverting X-axis





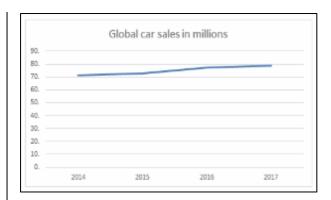
Explanation: Majority of the people have the tendency to assume that the time series data represented on the x-axis would always be increasing, as that is the conventional method. But if the x-axis is inverted, the scenario changes and the whole trend would be reversed. At the first glance, the misleading graph may convey the idea that the subscribers for Warcraft shows an increasing trend overall, which is not true.

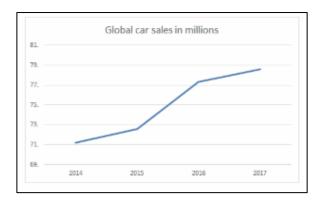
6. Worldwide Car Sales from 1990 to 2017

Bias Technique Used: Truncated Y-axis

Original Line Chart

Misleading Line Chart

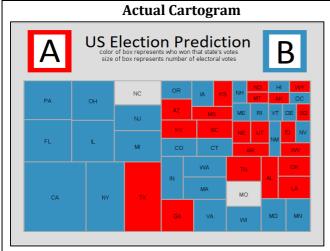


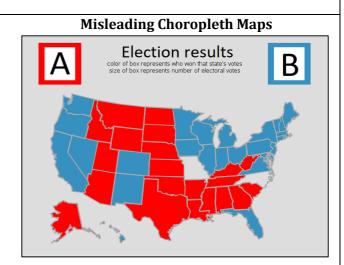


Explanation: When the Y-axis of the graph is truncated and is made to start from a custom value, the minimal growth trend of the sales can be portrayed as a major growth. As we can see from the accurate representation of the car sales data, the growth is very minimal and is almost stagnant towards the end. But by applying the biased technique, the user perceptions could be manipulated.

7. Election results in year 2008, A (Mitt Romney) and B (Barack Obama)

Bias Technique Used: Map Distribution

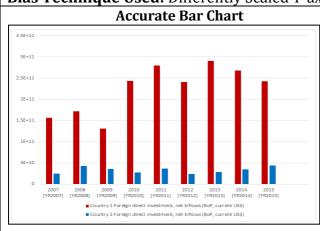


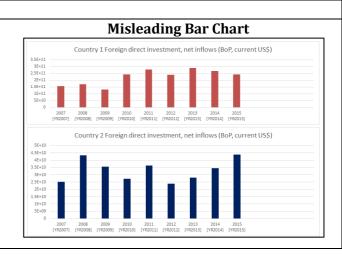


Explanation: The choropleth maps do not consider the population density and could be misleading for most people at the first glance rather than the cartogram which shows the perfect picture of the winner of 2008 US elections i.e. Barack Obama. The intensity of colors too provide a minor contribution to the magnitude of influence on viewer perceptions.

8. Comparison of Foreign Direct Investment between India & China

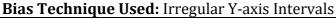
Bias Technique Used: Differently Scaled Y-axis

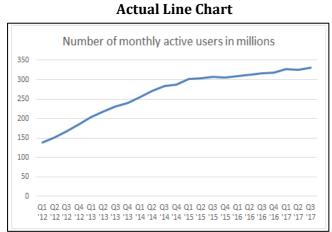


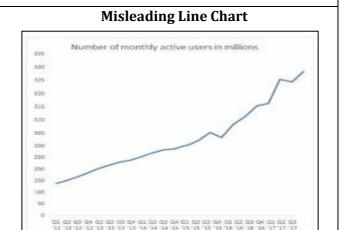


Explanation: In both the graphical representations the blue bars represent India and the red bars represent China. The users get the accurate perception only when the two bar charts are compared on the same scale. On the other hand, when the FDI investments in India and China are plotted spearately on different scales, the difference seems between them seem to be marginal for every year. However, that is not the case.

9. Number of monthly active users from 2012 to 2017



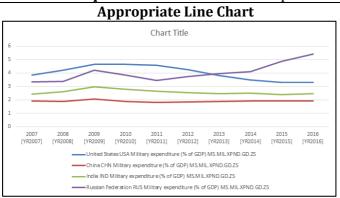


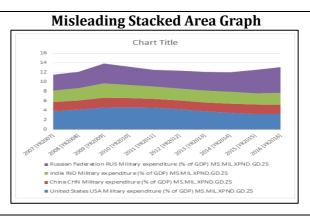


Explanation: Twitter saw a bullish growth in number of monthly active users from 2012 to 2015. But after 2015, the number of active users per month began to saturate, as we can infer from the actual representation. The misleading line graph having irregular y-axis intervals, possibly depicts that the growth in number of active monthly users continue to rise extensively, barring a few outliers.

10. Military Expenditure(% GDP) of different Countries from 2007 to 2016

Bias Technique Used: Stacked Area Graph





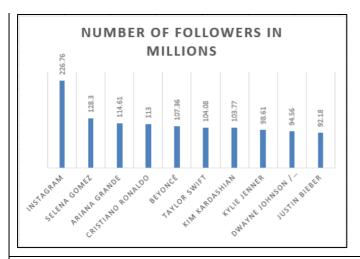
Explanation: Most of the participants tend to be misled by the stacked area graph representation of a scenario, as they misinterpret the curves to be indicating the quantitative measure. They sometimes forget the fact that in stacked area graphs, the curve of the area below serves as the baseline for the area at the top, whereas in the actual representation using the line chart all curves share the same baseline.

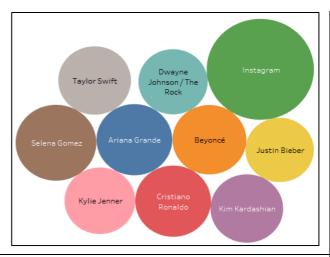
11. Instagram accounts with the most followers worldwide as of October 2017 (in millions)

Bias Technique Used: Bubble Charts

Appropriate Bar Graph

Misleading Bubble Charts





Explanation: Based on the misleading visualization, if a question is asked as "Do you think the followers for Christiano Ronaldo and Kim Kardashian in instagram are almost equal?", most of the participants would say 'Yes'. However, the difference of over 10 million followers as seen from the bar graph representation, was deflated by this visualization and made the bubbles look similar in size.

4 References

- [1] https://paragraft.wordpress.com/2008/06/03/the-chart-junk-of-steve-jobs/
- [2] https://www.mediamatters.org/blog/2012/07/31/dishonest-fox-chart-bush-tax-cut-edition/189046
- [3] Beattie, Vivien; Jones, Michael John (June 1, 1999). "Financial graphs: True and Fair?". Australian CPA. 69 (5): 42–44.
- [4] Beattie, Vivien; Jones, Michael John (1 September 1992). <u>"The Use and Abuse of Graphs in Annual Reports: Theoretical Framework and Empirical Study"</u>
- [5] Beattie, Vivien; Jones, Michael John (2002). "Measurement Distortion of Graphs in Corporate Reports: An Experimental Study"
- [6] Christensen, David S.; Albert Larkin (Spring 1992). "Criteria for High Integrity Graphics". Journal of Managerial Issues. <u>Pittsburg State University</u>. **4** (1): 130–153. <u>JSTOR</u> 40603924
- [7] Evaluation Part 2, COMP47490 Machine Learning Lecture Slides, Prof. Derek Greene
- [8] https://www.statista.com/
- [9] http://databank.worldbank.org
- [10] http://robslink.com/SAS/democd9/elect08.htm#map

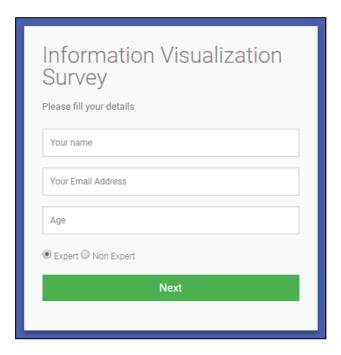
5 Appendix

5.1 Abbreviations

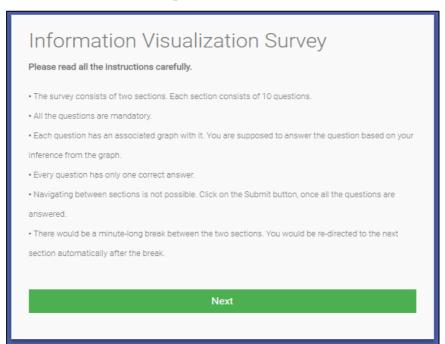
WWDC	Apple Worldwide Developer Conference
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristic

5.2 Participant Questionnaire

5.2.1 Registration Page

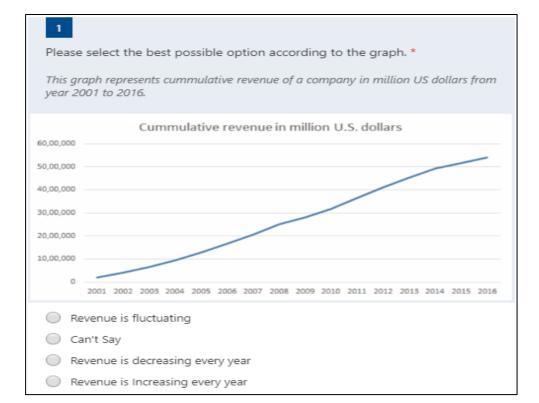


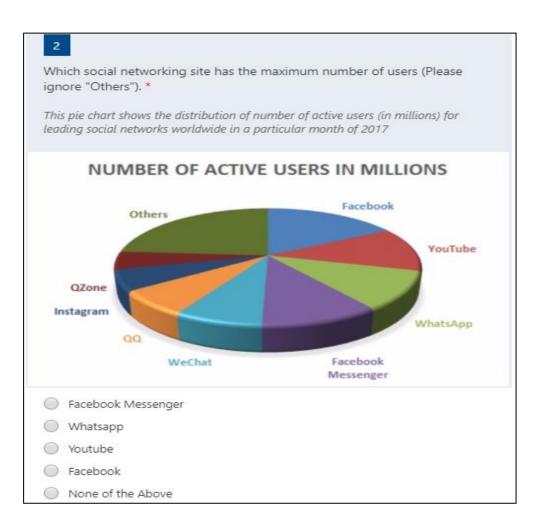
5.2.2 Instructions for Participants

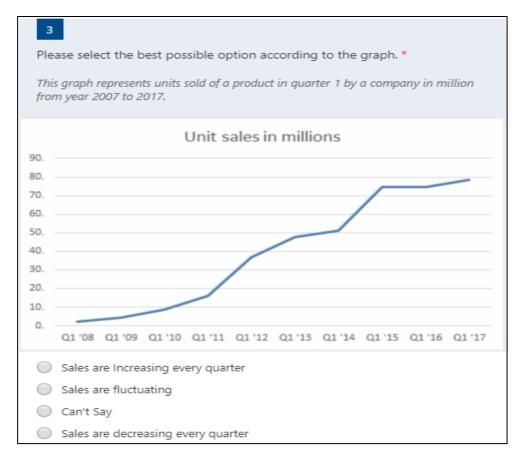


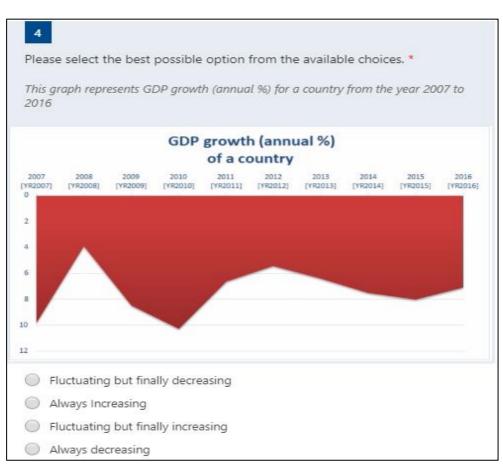
5.2.3 Section A - Biased Visualizations

Section A: Information Visualization General Survey





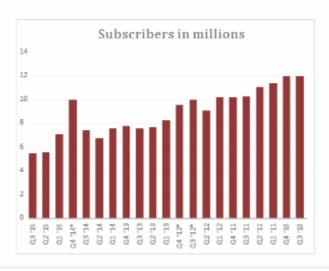




5

Please select the best possible option from the available choices. *

The graph represents number of subscribers of a game in millions in a quarter from the year 2010 to 2015

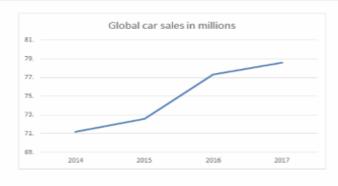


- Overall Increasing (Ignoring some random peaks)
- Overall Decreasing (Ignoring some random peaks)

6

Please select the best possible option from the available choices. *

This graph shows the yearly global car sales in millions from 2014 to 2017



- Sales are increasing drastically every year
- Sales are decreasing drastically every year
- Don't know
- Sales are decreasing but the decrease is marginal
- Sales are increasing but the increase is marginal

Which party seems to have won the election based on the graph shown.

*

This graph represents US election results for a particular year.

Election results

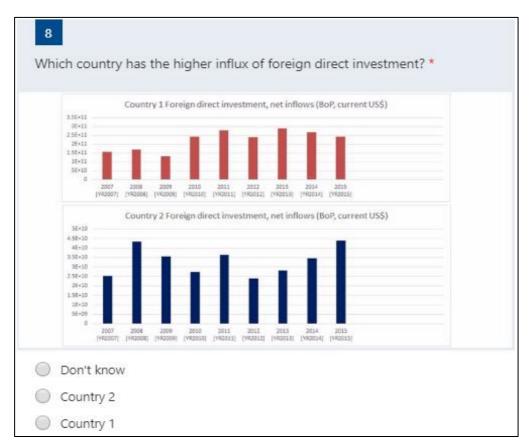
color of box represents who won that state's votes size of box represents number of electoral votes

B

It's a tie between Candidate A and Candidate B

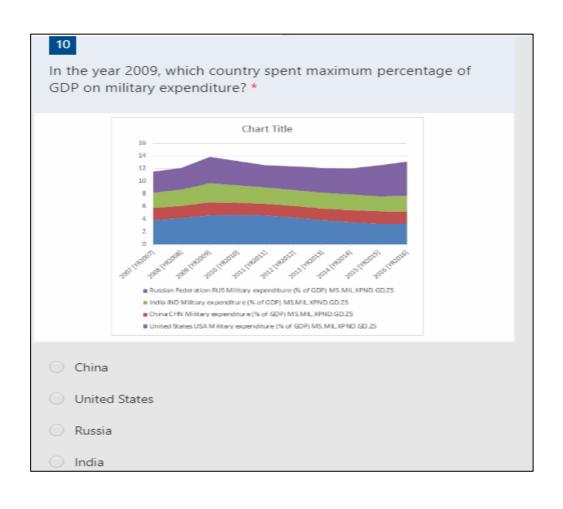
Candidate B Won

Candidate A Won



9 Please select the best possible choice from the available options. * The graph shows number of monthly active users on a popular social networking site in millions. Number of monthly active users in millions 330 325 320 315 310 305 500 250 200 150 100 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 04 01 02 03 12 12 12 13 13 13 13 13 14 14 14 14 15 15 15 15 15 16 16 16 16 16 17 17 17 17 Steady increase followed by stagnation Always decreasing Always increasing

Steady decrease followed by stagnation





5.2.4 Break Between Section A and Section B

00:57 seconds left

Please wait for the Section B.

5.2.5 Section B - Accurate Visual Representations

Section B: Information
Visualization General Survey

Please select the best possible option according to the graph. * This graph represents revenue of a company in million US dollars from year 2001 to 2016. Revenue in million U.S. dollars 5,00,000 4,50,000 4,00,000 3,50,000 3,00,000 2,50,000 2,00,000 1,50,000 1,00,000 50,000 Revenue is decreasing every year Revenue is fluctuating Revenue is Increasing every year Can't Say

Which social networking site has the maximum number of users (Please ignore "Others"). * This pie chart shows the distribution of number of active users (in millions) for leading social networks worldwide in a particular month of 2017 **NUMBER OF ACTIVE USERS IN MILLIONS** **Facebook** **Facebook** **PouTube** **WhatsApp** **PouTube** **PouTube**

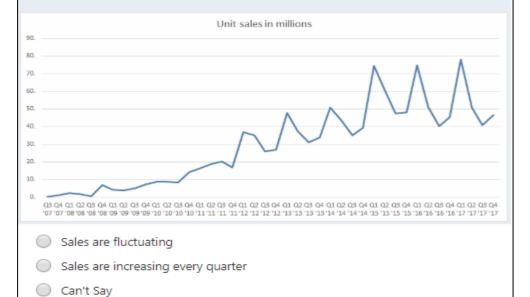
2

Youtube

ease s

Please select the best possible option according to the graph.*

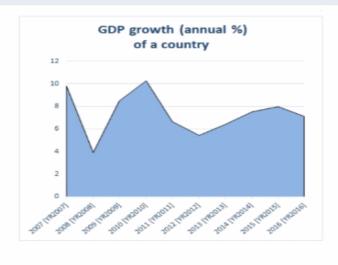
This graph represents units sold of a product quarterly by a company in million from year 2007 to 2017.



4

Please select the best possible option from the available choices. *

This graph represents GDP growth (annual %) for a country from the year 2007 to 2016



- GDP is fluctuating but eventually decreasing
- GDP is continuously decreasing

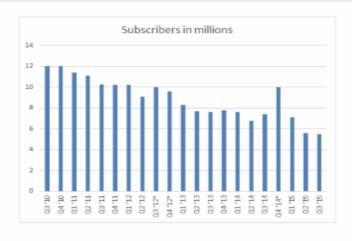
Sales are decreasing every quarter

- ODP is continuously increasing
- GDP is fluctuating but eventually increasing

5

Please select the best possible option from the available choices. *

The graph represents number of subscribers of a game in millions in a quarter from the year 2010 to 2015

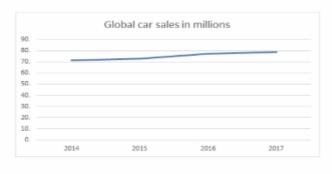


- Don't know
- Overall decrease in subscribers (ignoring some random peaks)
- Overall increase in subscribers (ignoring some random peaks)

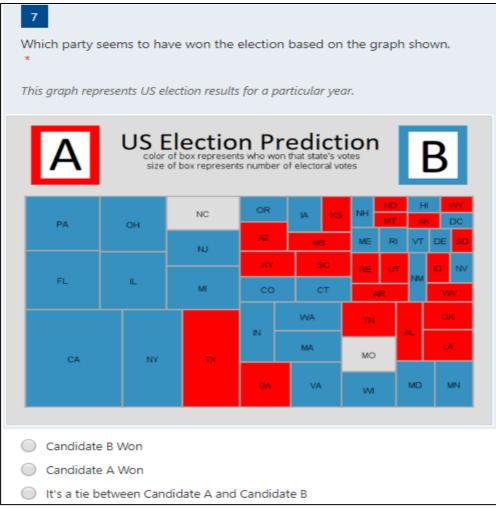
6

Please select the best possible option from the available choices. *

This graph shows the yearly global car sales in millions from 2014 to 2017



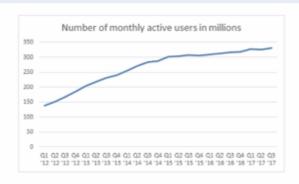
- Sales are increasing drastically every year
- Sales are decreasing drastically every year
- Oon't know
- Sales are decreasing but the decrease is marginal
- Sales are increasing but the increase is marginal



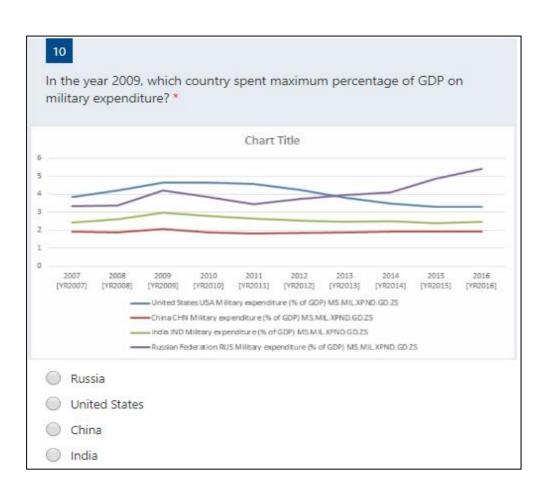


Please select the best possible choice from the available options. *

The graph shows number of monthly active users on a popular social networking site in millions.

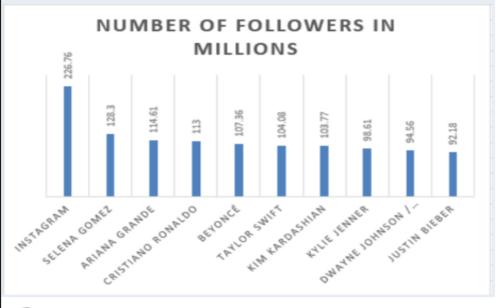


- Steady increase followed by stagnation
- Steady decrease followed by stagnation
- Ontinuous decrease in no. of users
- Continuous increase in no. of users





From the below graph, do you think both Christiano Ronaldo and Kim Kardashian have approximately the same number of followers on Instagram?



Yes

Can't Say

O No