|  |
| --- |
| DATA VISUALIZATION PROJECT  Part 1: Features in Tableau  Part 2: Exploration of a Dataset |
| **Yash Sinojia**  **A00268852**  **M.Sc. Data Analytics** |





**Faculty of Business and Hospitality**

**2019-2020 Academic Year**

**Semester 2**

**Subject: Data Visualization**

**Student Name: Yash Sinojia**

**Student ID: A00268852**

**Course: Master of Science in Data Analytics**

**Date: 27-Apr-2020**

**Title: Data Visualization Project**

**Part 1: Features in Tableau**

**Part 2: Exploration of a Dataset**

**TABLE OF CONTENTS**

[**1. INTRODUCTION 4**](#_Toc38846214)

[**2. FREATURES OF TABLEAU 5**](#_Toc38846215)

[**2.1. Feature 1: Filters and Markers 5**](#_Toc38846216)

[**2.1.1. Explanation 5**](#_Toc38846217)

[**2.1.2. Example 6**](#_Toc38846218)

[**2.2. Feature 2: Calculated Fields 12**](#_Toc38846219)

[**2.2.1. Explanation 12**](#_Toc38846220)

[**2.2.2. Example 12**](#_Toc38846221)

[**2.3. Feature 3: Dashboard Functionalities 14**](#_Toc38846222)

[**2.3.1. Explanation 14**](#_Toc38846223)

[**2.3.2. Example 14**](#_Toc38846224)

[**3. EXPLORATION OF A DATASET 17**](#_Toc38846225)

[**3.1. The Dataset 17**](#_Toc38846226)

[**3.2. Exploratory Analysis 18**](#_Toc38846227)

[**3.3. Questions 23**](#_Toc38846228)

[**3.4. Visualized Answers 24**](#_Toc38846229)

[**4. REFERENCES 30**](#_Toc38846230)

# INTRODUCTION

Tableau is a business intelligence and data visualization tool that renders users to make sense of their data through interactive charts, graphs, and diagrams. It is a tool definitively used for the Exploration and Reporting stage of ‘The Data Science Workflow’.

Tableau is easy to use tool with its drag and drop functionalities and automatic suggestions for charts and visualizations. Tableau identifies a dataset and all the hierarchies of categories associated with it. The default visualizations generated by Tableau follows all the data visualization principles. Hence, it is a quite professional tool for business intelligence.

This project is divided into 2 parts. Part 1 revolves on exploring three essential features of Tableau as a Tutorial, where these three features would be demonstrated by an example and backed up with a corresponding explanation. Part 2 revolves around exploring a dataset. This analysis would be a basis in forming five questions and their corresponding insightful answers visualized as dashboards and stories.

# FREATURES OF TABLEAU

To explore the three features of Tableau, the Tableau 2019.4 inbuilt dataset ‘World Indicators’ has been utilized (*Indicators | Data*, no date). Each feature would be introduced briefly, which would be backed up by an explanation and demonstrated by a corresponding example as described by several worksheets. In this series of feature tutorials, the ‘Show Me’ worksheet functionality has been utilized to pick one among all the suggested visualization that Tableau provides.

## 2.1. Feature 1: Filters and Markers

This section explores the automated filters and markers provided by Tableau worksheet once the structure of a visualization is provided.

### **2.1.1. Explanation**

Tableau provides a Filters shelf to filter out the graphical view. There are four kinds of filters in Tableau:

1. **General Filter:**

Provides general filtering on the basis of categories or hierarchical sub-categories.

1. **Wildcard Filter:**

Provides filtering on the basis of four wild cards:

1. Contains
2. Starts with
3. Ends with
4. Exactly matches
5. **Conditional Filter:**

Provides filtering adhering to some conditions, two options are provided:

1. By Field
2. By Formula

1. **Top Tab Filter:**

Provides filtering adhering to some conditions by field or by formula and also provides an additional optionality to pick top or bottom ‘n’ categories under it.

Tableau provides a mark section on the worksheets which adds certain visual characteristics on a visualization. Six distinct options are provided:

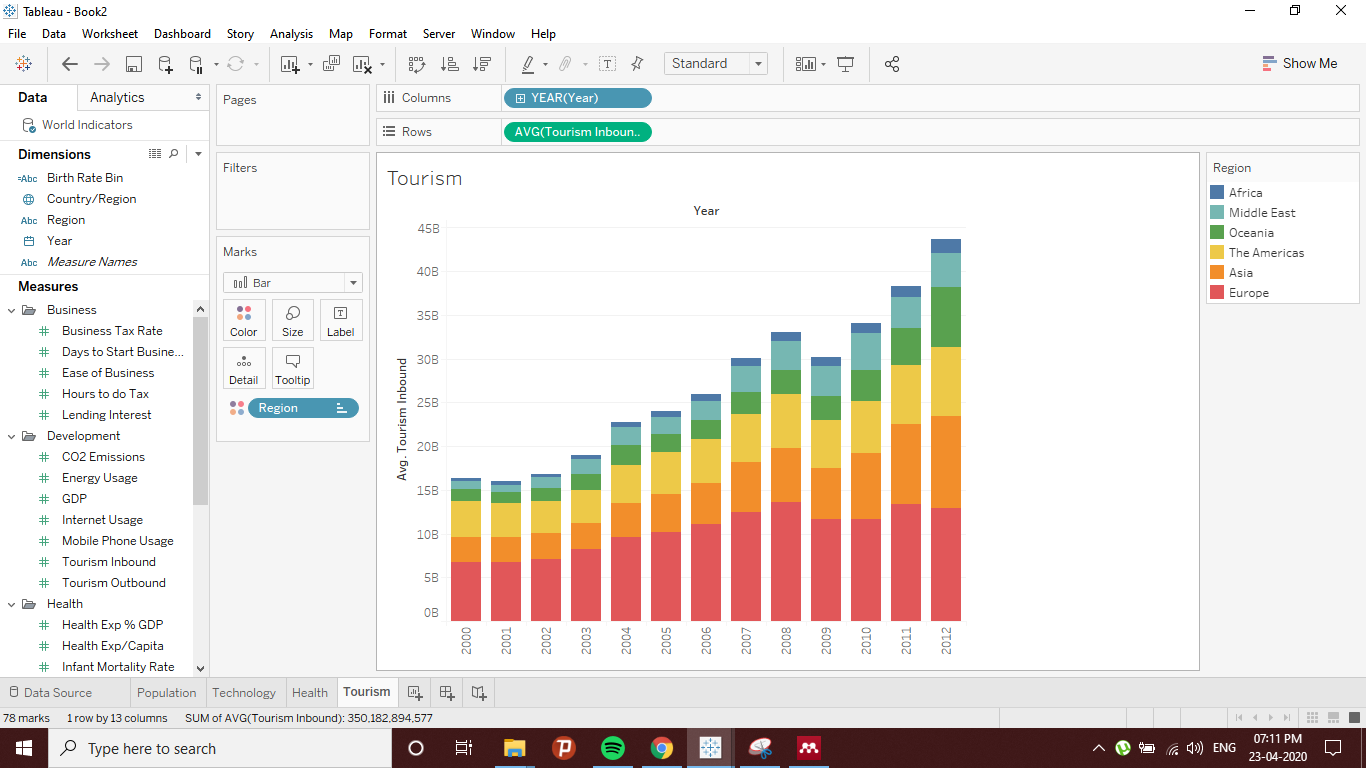
1. **Color**
2. **Size**
3. **Label**
4. **Detail**
5. **Tooltip**
6. **Shape**

### **2.1.2. Example**

To demonstrate the above functionalities 5 different worksheets are used to describe each individual filter example and one to describe the markers. For better demonstration, World Indicators dataset has been used.

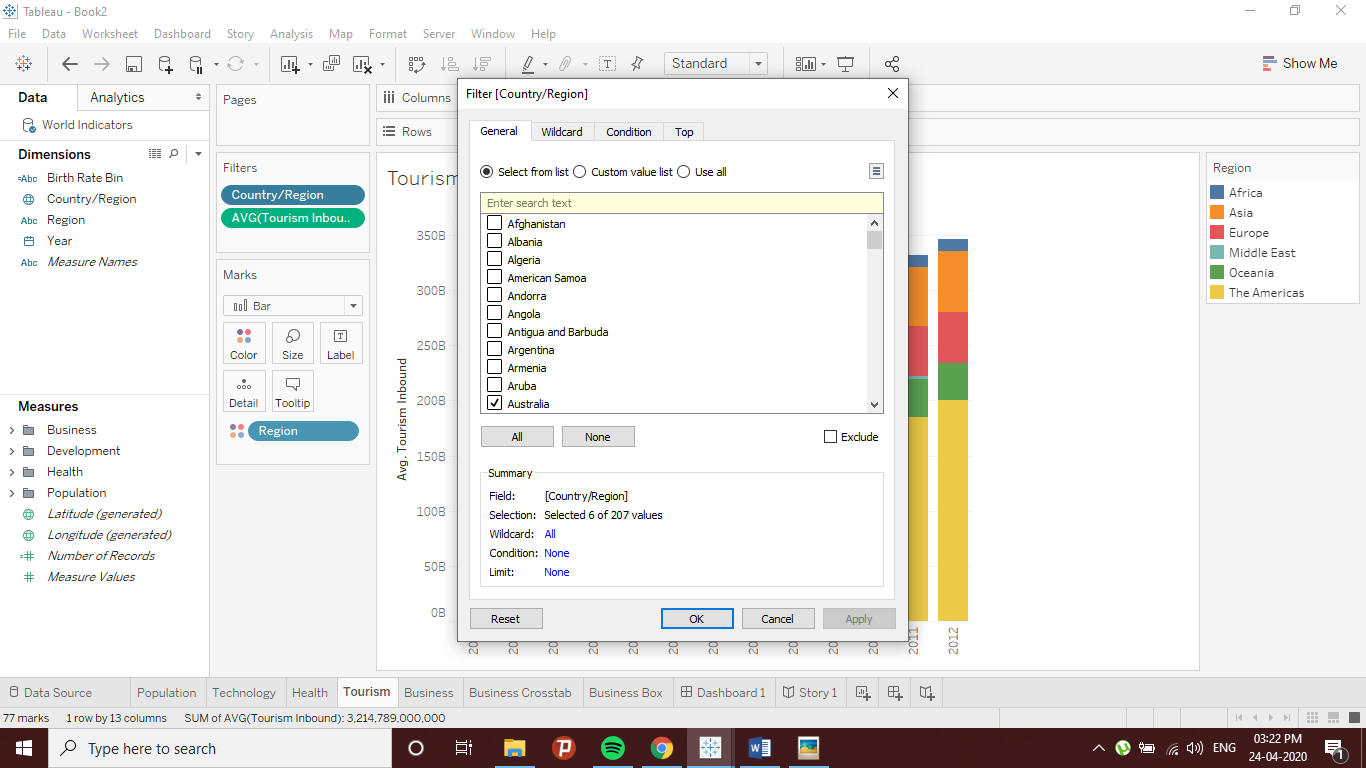
1. **General Filter**

Here, ‘Year’ is picked as dimension on columns and Average of ‘Tourism Inbound’ as a measure on rows. To add an extra layer on it, the dimension ‘Region’ is picked as colored category. Then, from the ‘Show Me’ section, a bar plot is picked that best describes this scenario (*Tableau Tutorial - Tutorialspoint*, no date).



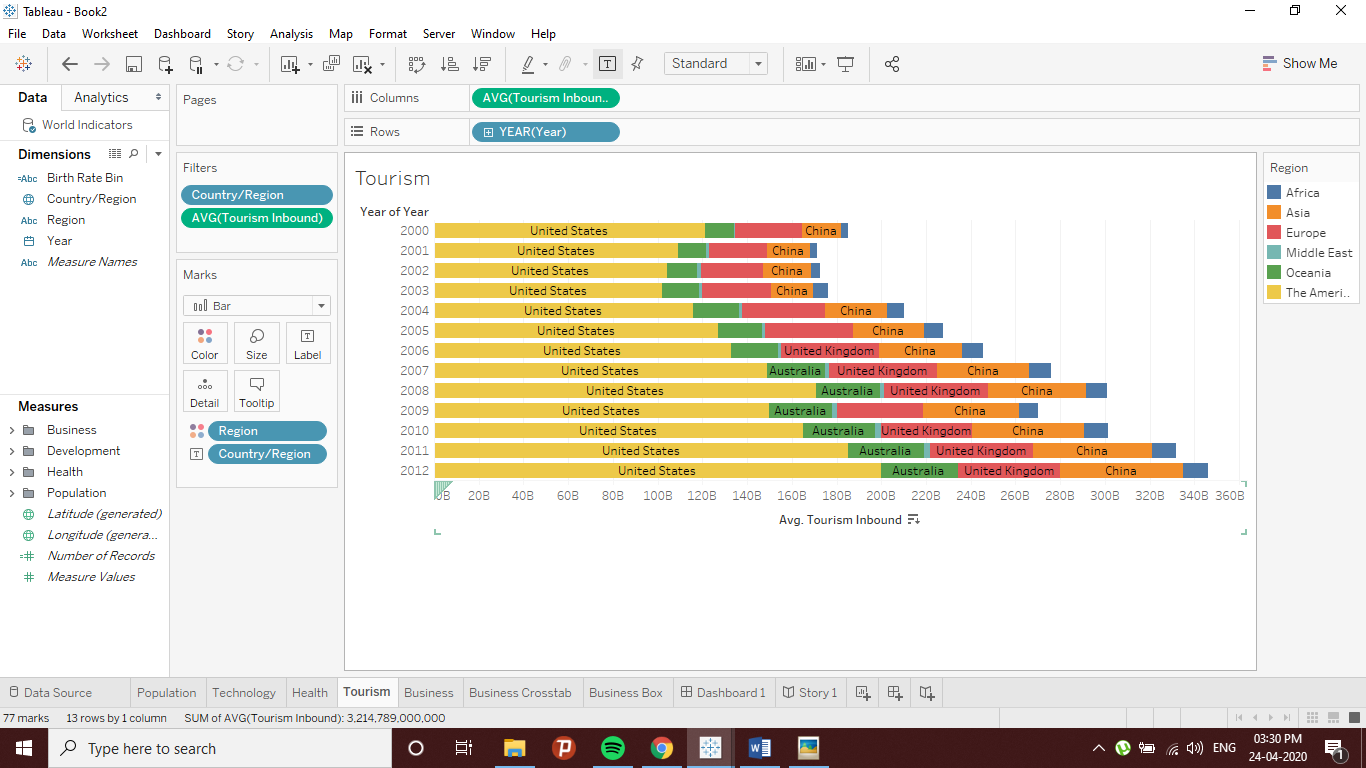
**Fig 1: General Filter Example**

Now, ‘Country’ from dimensions is dragged and dropped on Filters section.



**Fig 2: Filtering out Values**

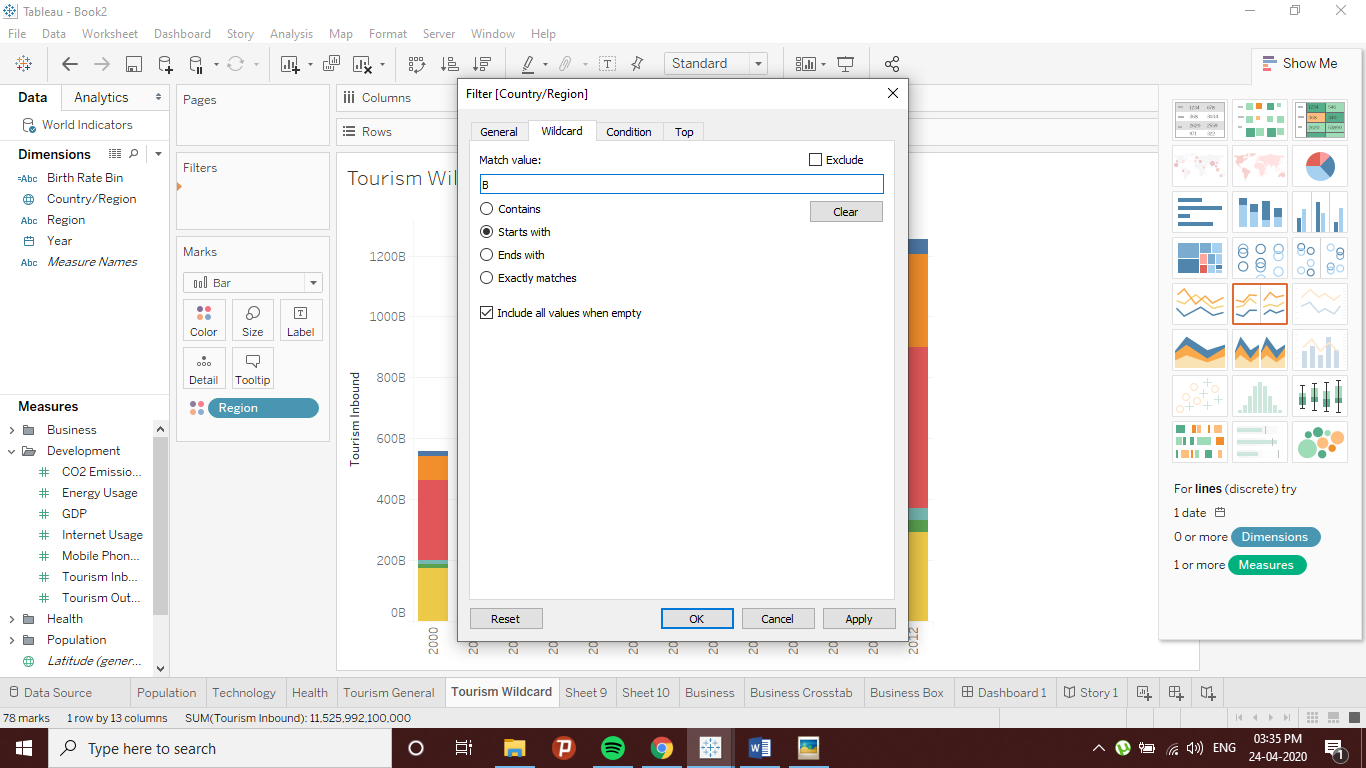
Here, a country from each region is picked randomly as follows: South Africa, Iran, Australia, United States, China and Germany respectively. The results on the visualization can be seen more clearly by adding labels to ‘Country’ and reversing the axes by using Transpose option.



**Fig 3: Transposing axes**

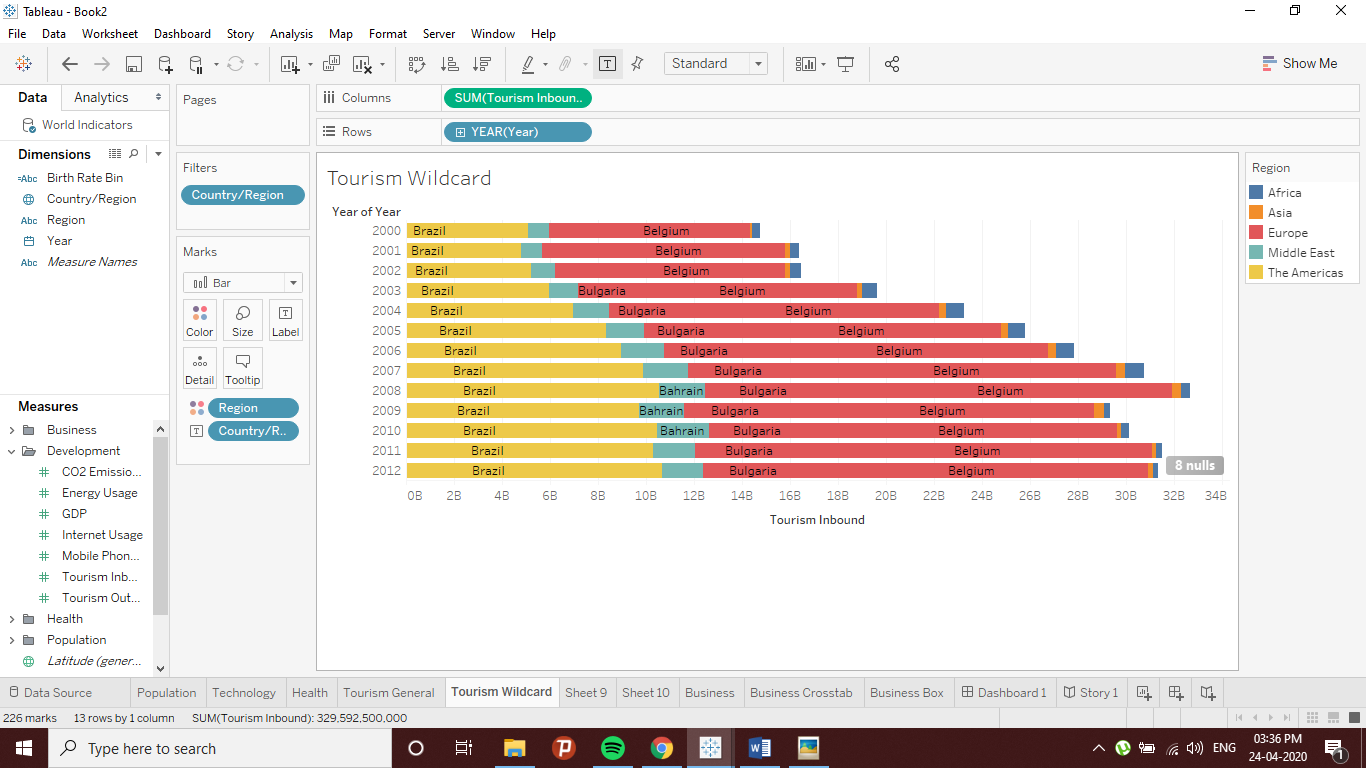
1. **Wildcard Filter**

Same variables are used for all the other filters. Again, dragging ‘Country’ on Filters field. Now, filtering by character ‘Starts with’ as a wildcard and character ‘B’ as match value.



**Fig 4: Wildcard Filter Example**

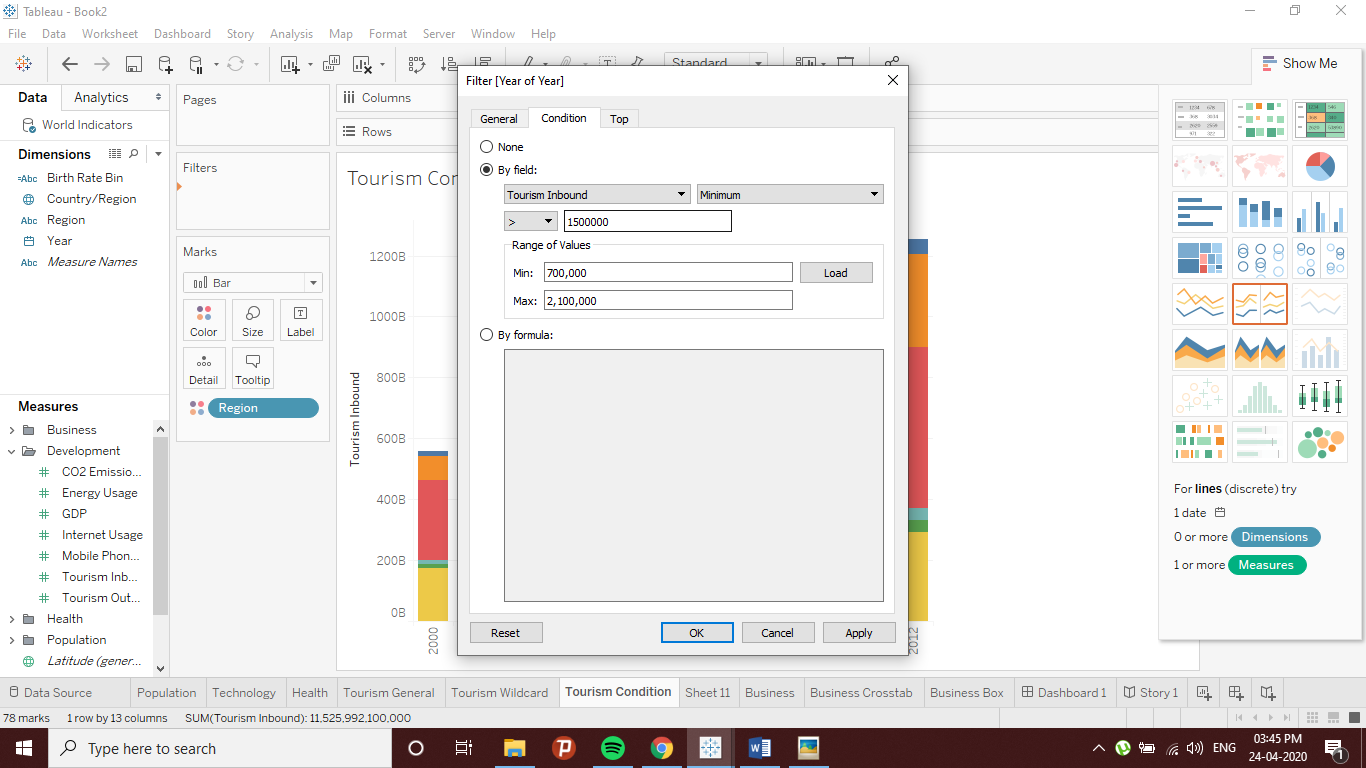
The resulting visualization labelled by ‘Country’ and transposed axes is as follows:



**Fig 5: Transposing axes**

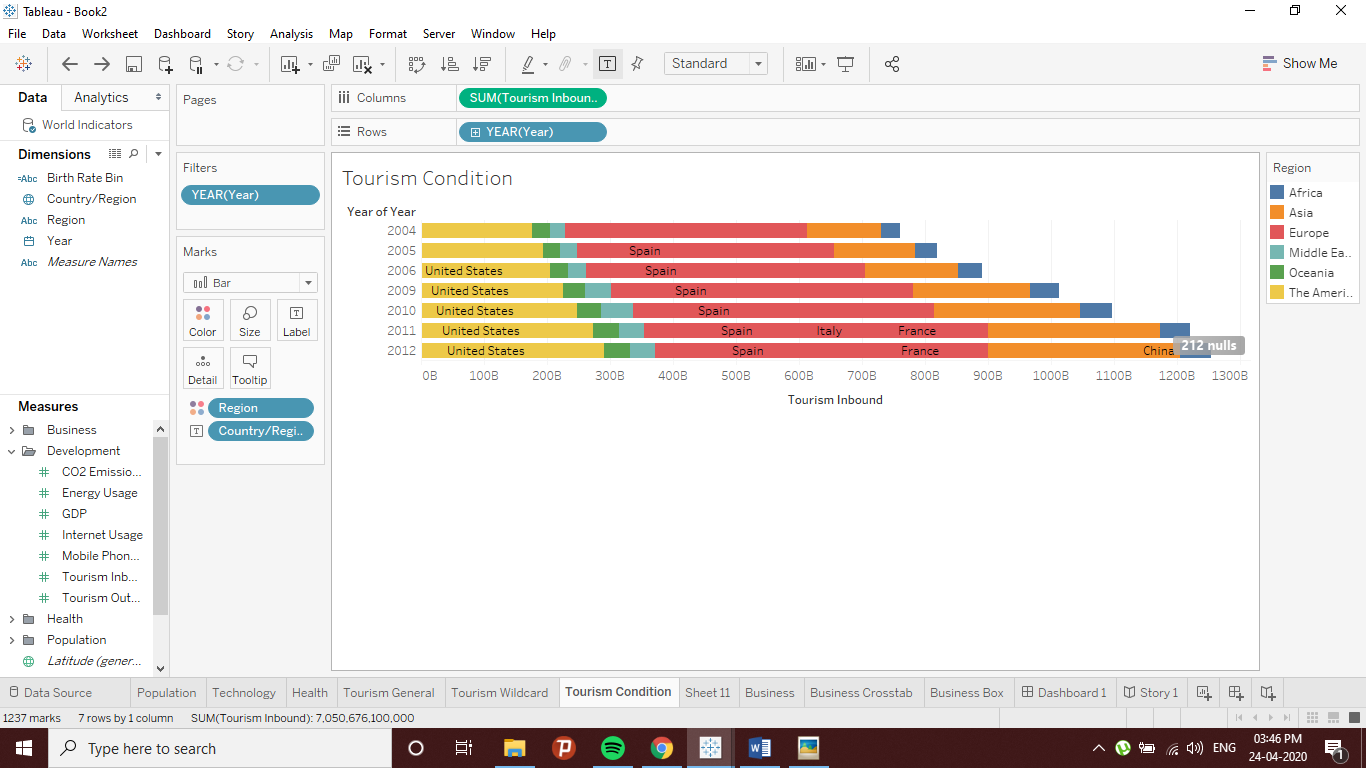
1. **Conditional Filter**

For this case, due to compatibility, ‘Year’ is being used as the Filter variable. The condition field here is ‘Tourism Inbound’ measure, the condition being the minimum > 1.5M.



**Fig 6: Conditional Filter Example**

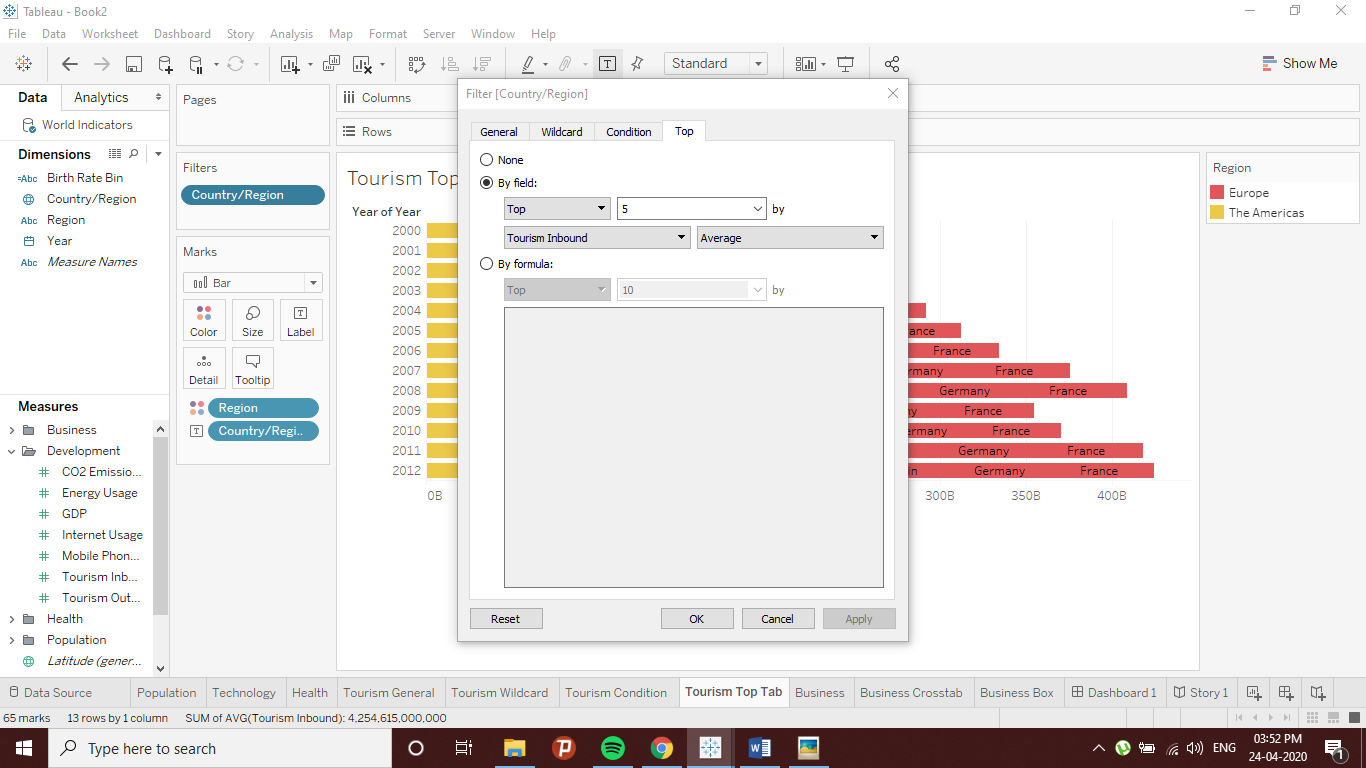
Again, labelling by ‘Country’ and transposing the axes for better view.



**Fig 7: Transposing axes**

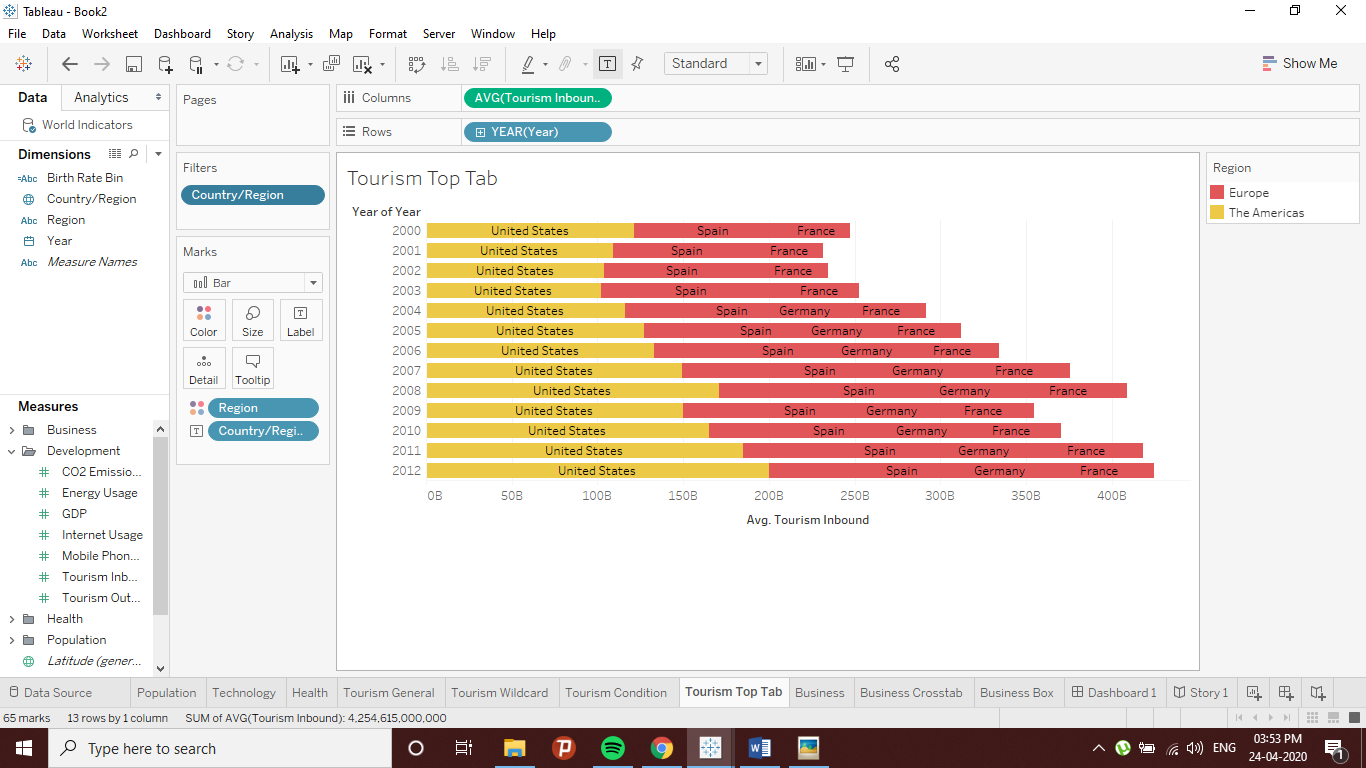
1. **Top Tab Filter**

Here, we refine ‘Country’ by displaying Top 5 entries on the basis of Average of ‘Tourism Inbound’.



**Fig 8: Top Tab Filter Example**

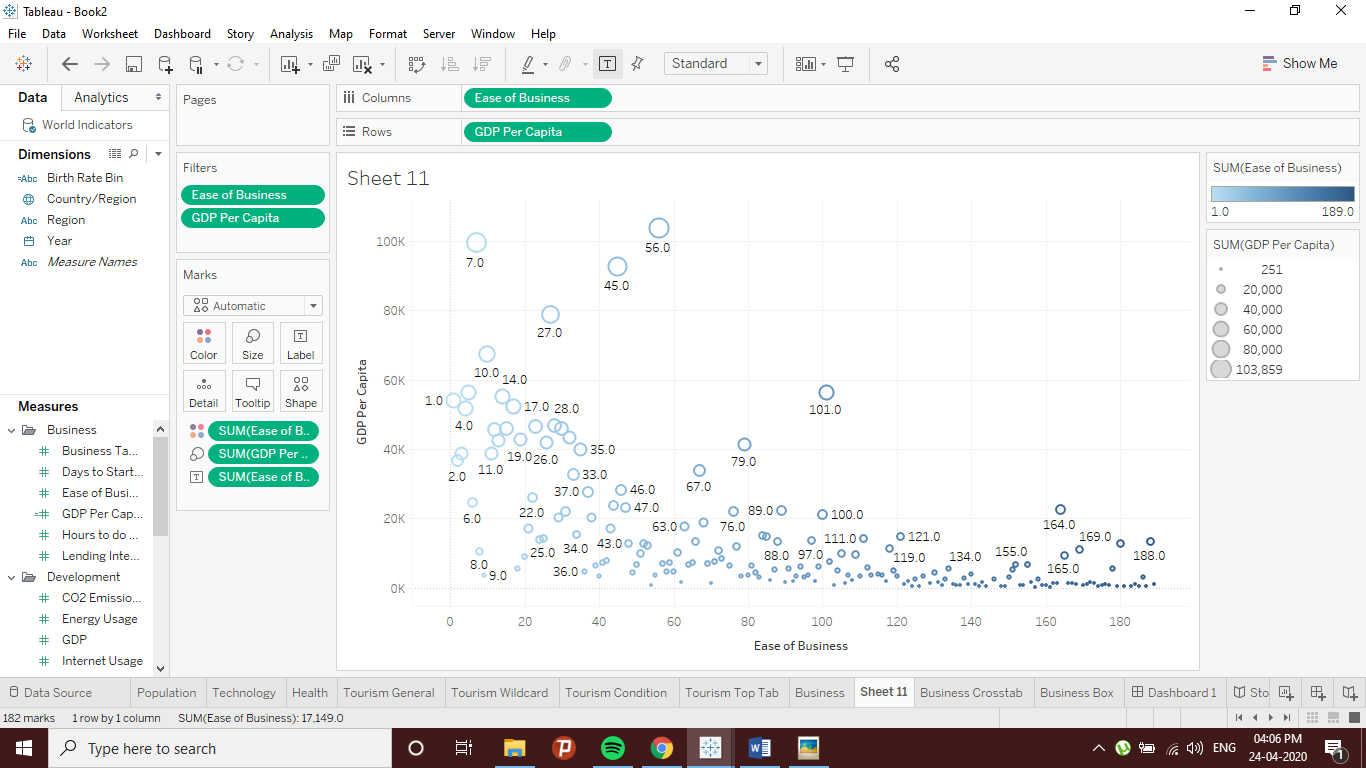
Labelling ‘Country’ and taking the transpose of axes.



**Fig 9: Transposing axes**

1. **Markers**

In this case, the measures ‘Ease of Business’ and ‘GDP Per Capita’ after converting them to dimensions are plotted on axes with a Scatterplot. To add visual layers, the sum of ‘Ease of Business’ is color coded and labelled while the sum of ‘GDP Per Capita’ has been given size marks as the range and labels provided below:



**Fig 10: Various Markers**

## 2.2. Feature 2: Calculated Fields

### **2.2.1. Explanation**

Other than Drag and Drop functionalities, Tableau also offers conditional looping functions like programming languages. Here, calculated fields are used to assign certain new categorical labels i.e. If-Else statements. This feature can be tested on bar plot by assigning certain bin conditions (*Tableau Tutorial - Tutorialspoint*, no date).

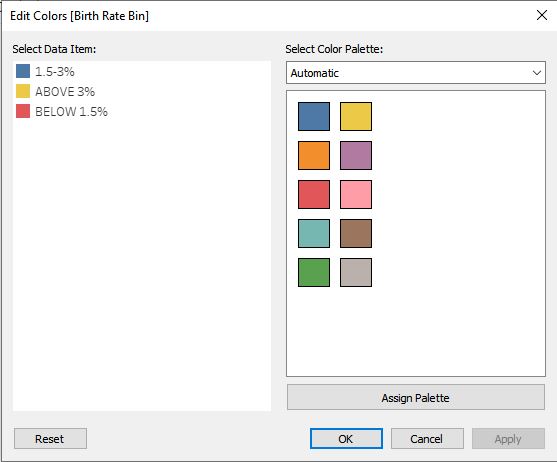
### **2.2.2. Example**

In Tableau visualizations, if longitudinal values are assigned as columns and latitudinal values as rows, they are automatically plotted on a world map. Using population data with world map plot, a calculated field called ‘Birth Rate Bin’ is made. An If-Else conditional statement is being assigned on ‘Birth Rate’ measure to make three new categorical bins as follows:



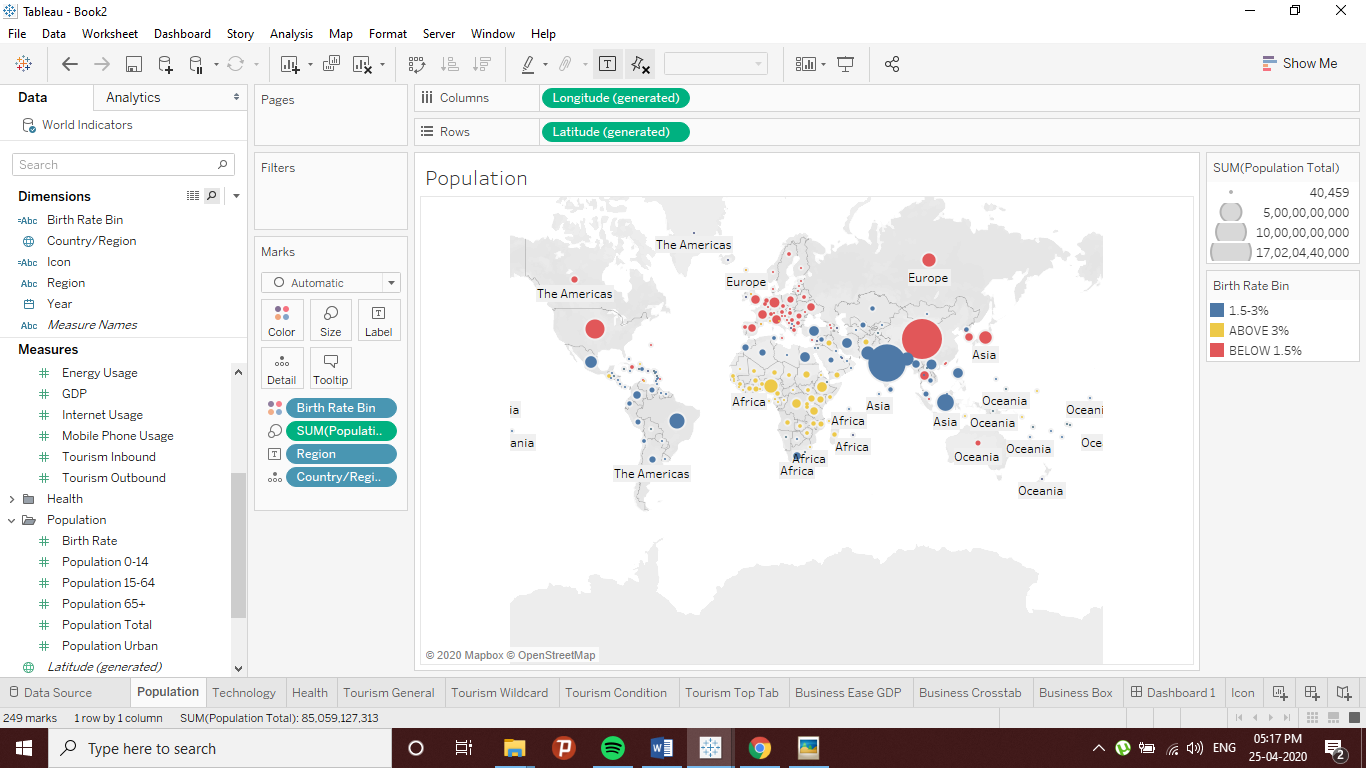
**Fig 11: A Calculated Field**

With options, a different color palette is selected to make more sense within these ordinal bins.



**Fig 12: Color Palette**

Other markers i.e. size of the sum of ‘Population Total’, labelling each Region and adding details on each ‘Country’ are used to add extra layers on map.



**Fig 13: Latitudinal - Longitudinal Plot**

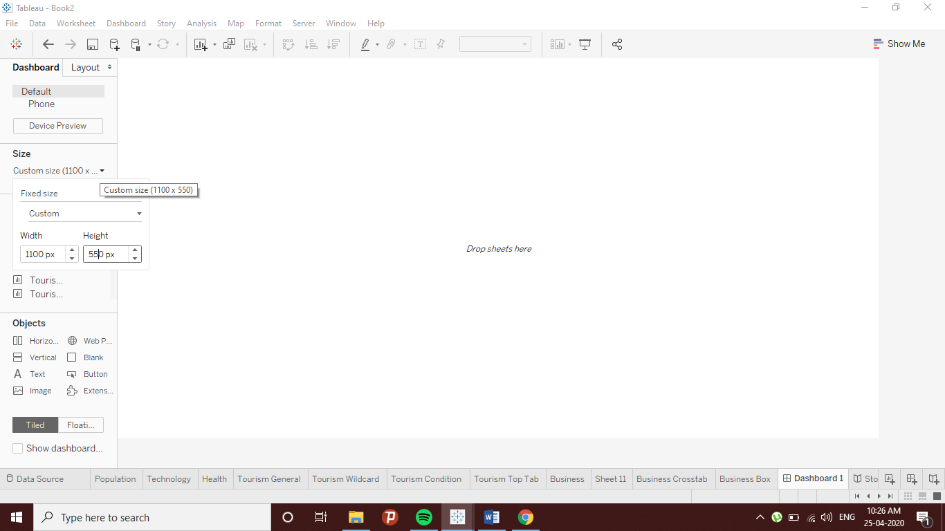
## 2.3. Feature 3: Dashboard Functionalities

### **2.3.1. Explanation**

Tableau provides Dashboards, where the visualizations generated from a set of Worksheets can be summarized. It also adds interactivity within the group of visualizations. Furthermore, it is easier to understand each individual sections of the workflow and interpret the message more clearly and lucidly.

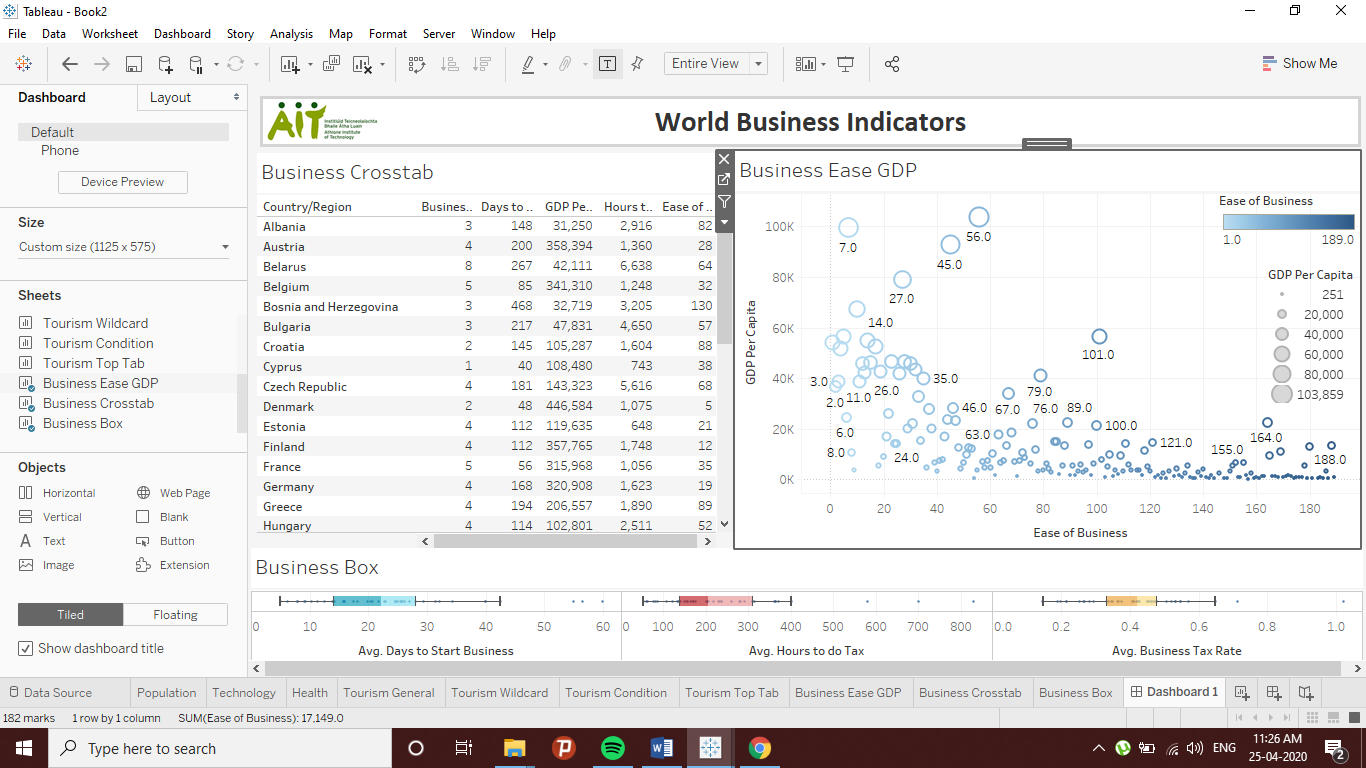
### **2.3.2. Example**

A new dashboard has a canvas to combine multiple worksheets. This canvas can be resized and colors can be added on layout as follows:



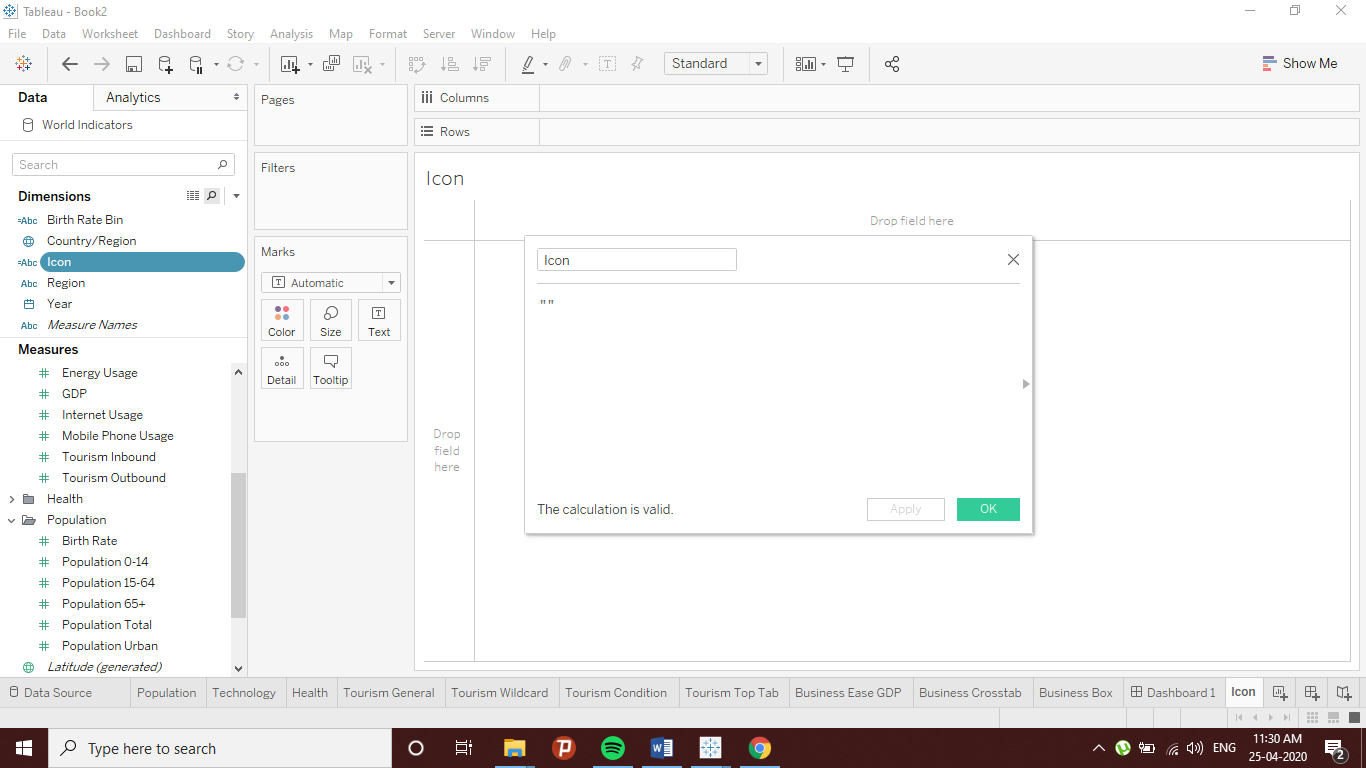
**Fig 14: Dashboard Canvas**

Professionally, the dashboard is divided into two sections: a header area and a plot area. On header area a heading and a logo is described. On plot area, three worksheets Business Crosstab, Business Ease GDP (Scatterplot) and Business Box (3 boxplots) are picked to display a combined analysis on World Business Indicators. Furthermore, the category area is dragged into Business Ease GDP as it is relevant only for that plot and not the others. Dashboards provides headings for each worksheet plots as default (*Tableau Dashboard Tips [Top 10 Tableau Dashboard Design Tips] - YouTube*, no date).



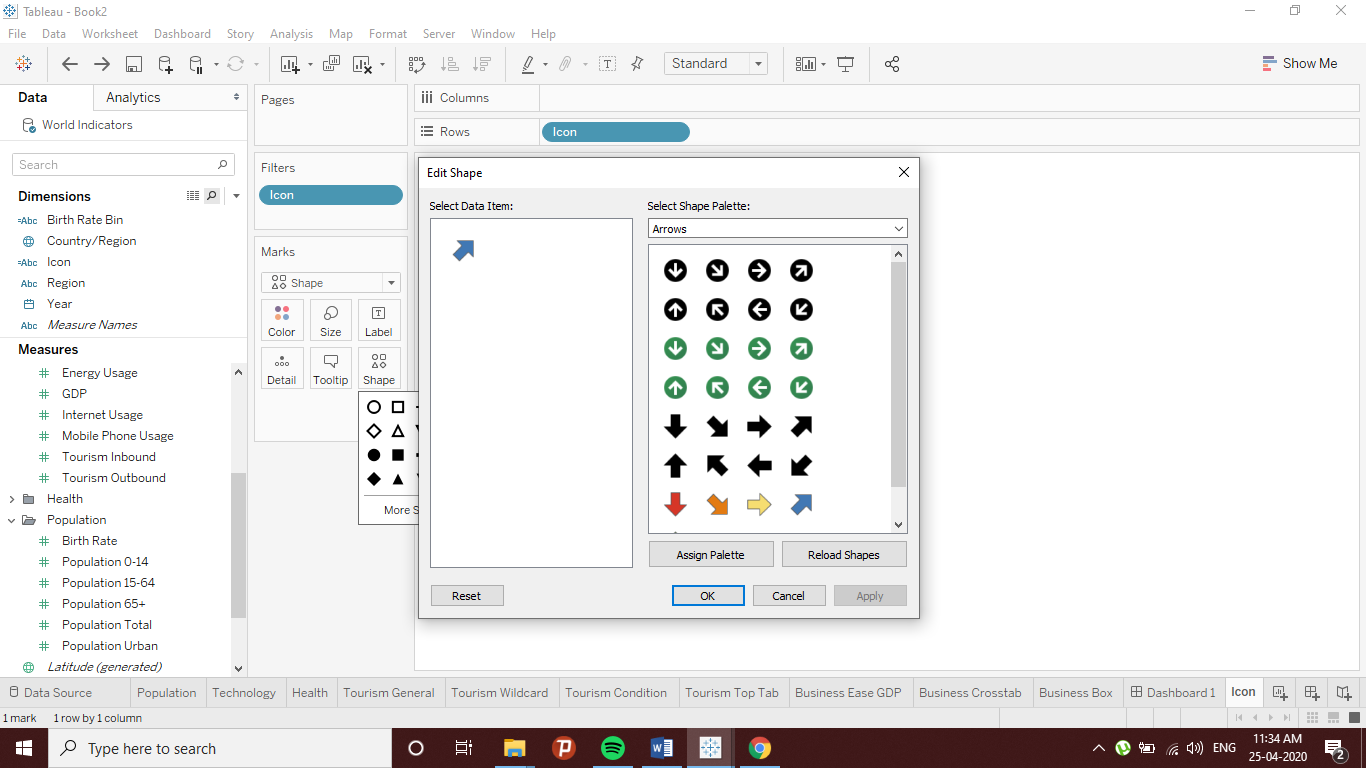
**Fig 15: Business Dashboard**

To make it more interactive, a logo is designed to link to other sources of information. This logo has to be first created in a separate worksheet as a calculated column. For now, a random double quote is assigned to make it a valid calculated field.



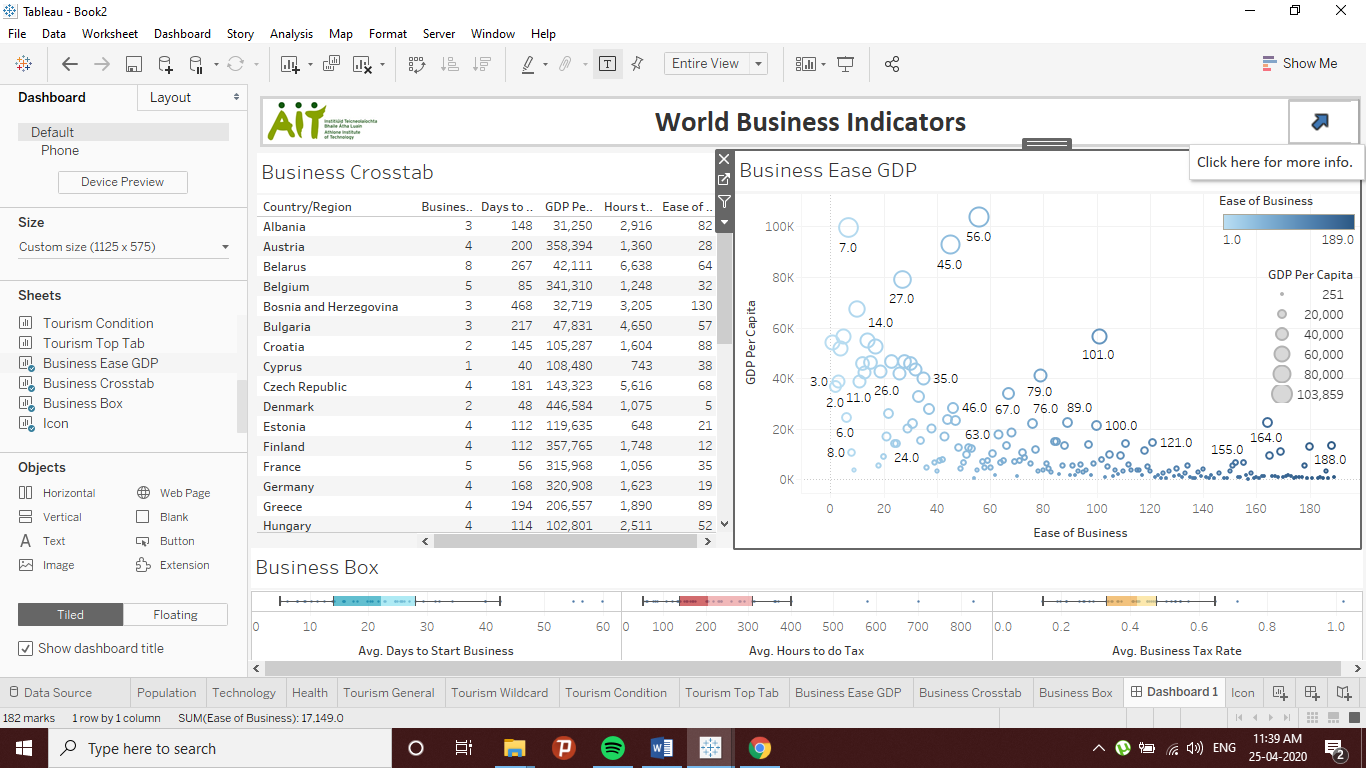
**Fig 16: A Calculated Field**

Then that value can be converted to a graphic by assigning it a shape palette in shape marker field. Here, a blue arrow is selected as an indication for more information. Furthermore, a tooltip has been also added on the icon to display out a short message whenever the mouse pointer hoovers on it.



**Fig 17: Shape Palette**

Now, the icon is ready to be placed on the header of the business dashboard. When hoovered by the pointer, it displays a short message on screen as shown below:



**Fig 18: Interactive Dashboard**

# EXPLORATION OF A DATASET

## 3.1. The Dataset

For this part, a Kaggle dataset called ‘Goodreads-books’ having adequate numbers of both categorical and numerical variables has been shortlisted.

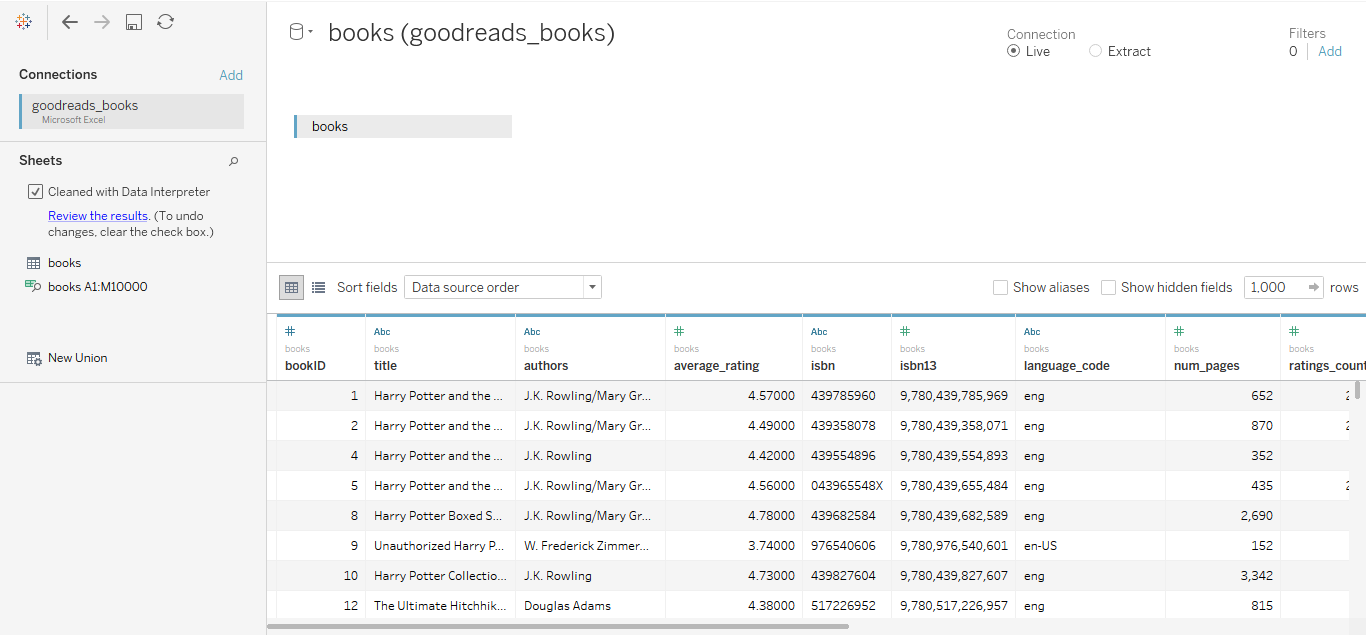
This dataset was entirely scraped via Goodreads API. Goodreads is an internet database of books as well as a social media for book lovers. Goodreads provides a simple interface to scrape their database (*Goodreads-books | Kaggle*, no date).

This dataset is stored as ‘goodreads\_books.xslx’ file and has been imported into Tableau as an Excel Worksheet.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Columns** | **Row Values** | **Description** |
| 1. | bookID | Numerical  (1 – 45.6k) | A unique Identification number for each book. |
| 2. | title | Categorical | The name under which the book was published. |
| 3. | authors | Categorical | Names of the authors of the book. Multiple authors are delimited with -. |
| 4. | average\_rating | Numerical  (0 – 5) | The average rating of the book received in total. |
| 5. | isbn | Numerical  (Unique Values) | Another unique number to identify the book, the International Standard Book Number. |
| 6. | Isbn13 | Numerical  (Unique Values) | A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN. |
| 7. | language\_code | Categorical | Helps understand what is the primary language of the book. For instance, eng is standard for English. |
| 8. | num\_pages | Numerical | Number of pages the book contains. |
| 9. | ratings\_count | Numerical | Total number of ratings the book received. |
| 10. | text\_reviews\_count | Numerical | Total number of written text reviews the book received. |
| 11. | publication\_date | Date | Date when the book was first published. |
| 12. | publisher | Categorical | The name of the publisher. |

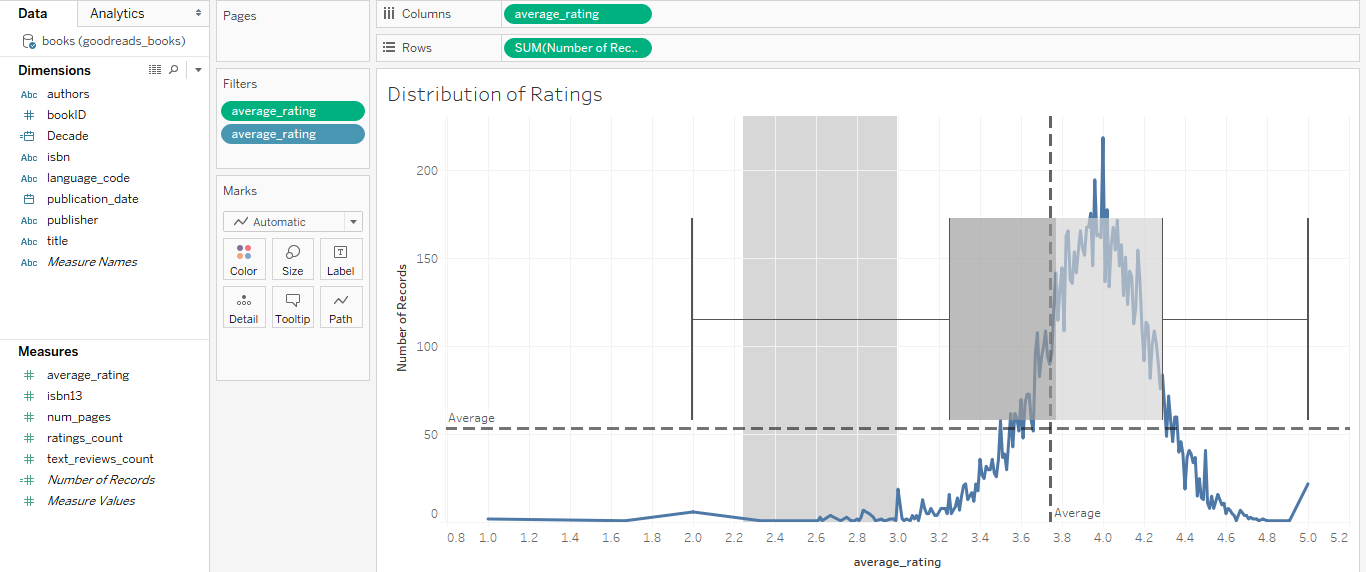
**Source:** (*Goodreads-books | Kaggle*, no date)

## 3.2. Exploratory Analysis



**Fig 19: Data Source**

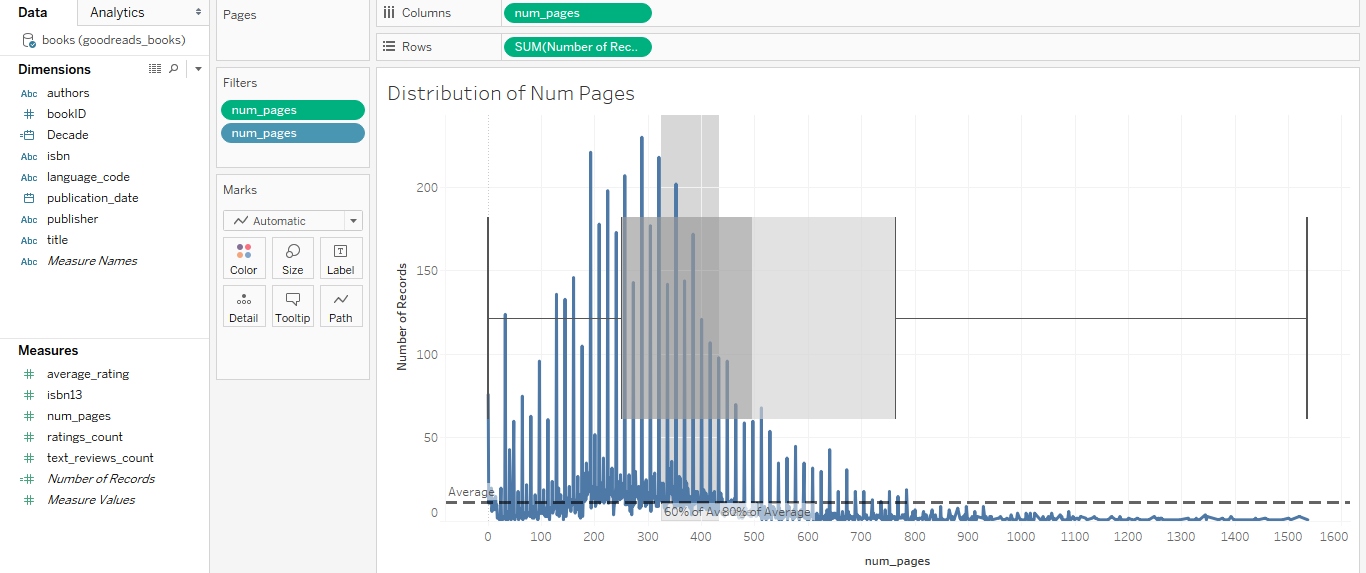
The .xlsx file has been connected with Tableau, the connection is kept Live as in default. Data Interpreter has been used for a quick default cleanup. To start with the analysis, it is essential to go through the distributions of measures. Starting with distribution of ratings.



**Fig 20: Distribution of Ratings Worksheet**

Plot describes the distribution of ‘average\_ratings’ with number of records. To understand this distribution more clearly, Average lines are added from Analytics section provided by Tableau worksheet. From the same section, a translucent boxplot has been juxtaposed upon the center of the distribution. Following it, 3.746 is the average rating value and the distribution a bit right skewed. There has been null values and some extreme outliers which have been excluded and filtered out for better view of distribution. This distribution is plotted on a line plot as the ‘average\_rating’ has been converted to a dimension with sum of ‘Number of Records’ as a measure. Number of Records is provided as a default calculated field by Tableau.

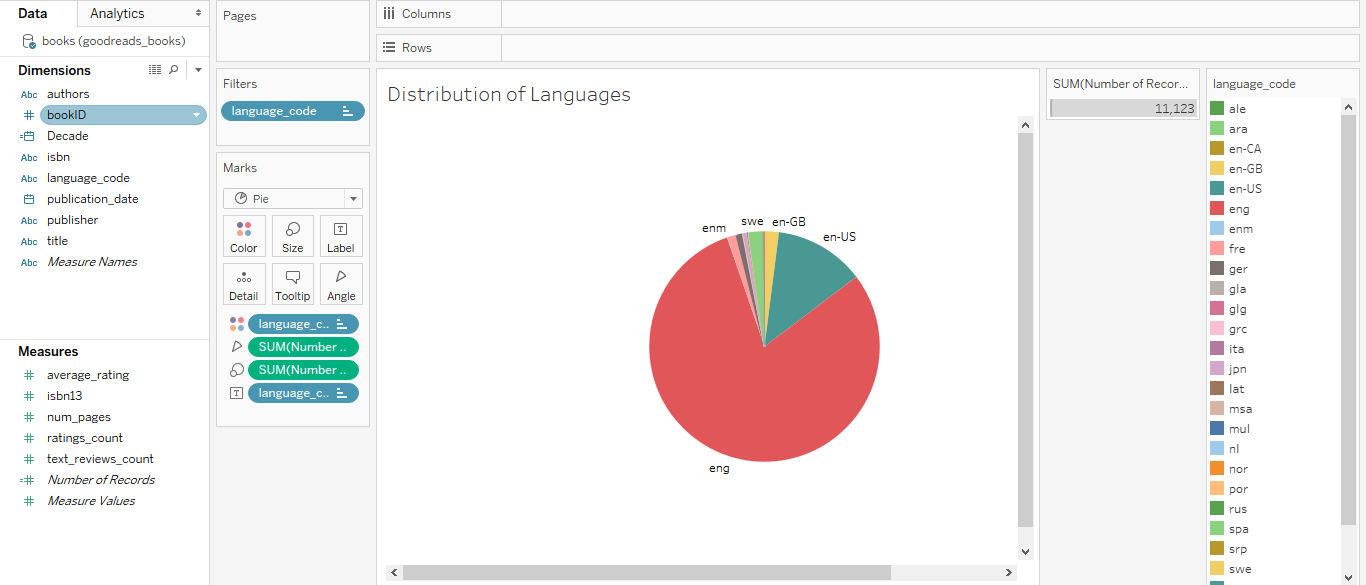
The same process has been done on ‘num\_pages’ measure.



**Fig 21: Distribution of Num Pages Worksheet**

Here, the distribution is quite left-skewed and a significant number of outliers from the right side (higher values) have been excluded and filtered out to get a better view of the distribution. The average number of pages are 496. This distribution is plotted on a line plot as the ‘num\_pages’ has been converted to a dimension with sum of ‘Number of Records’ as a measure.

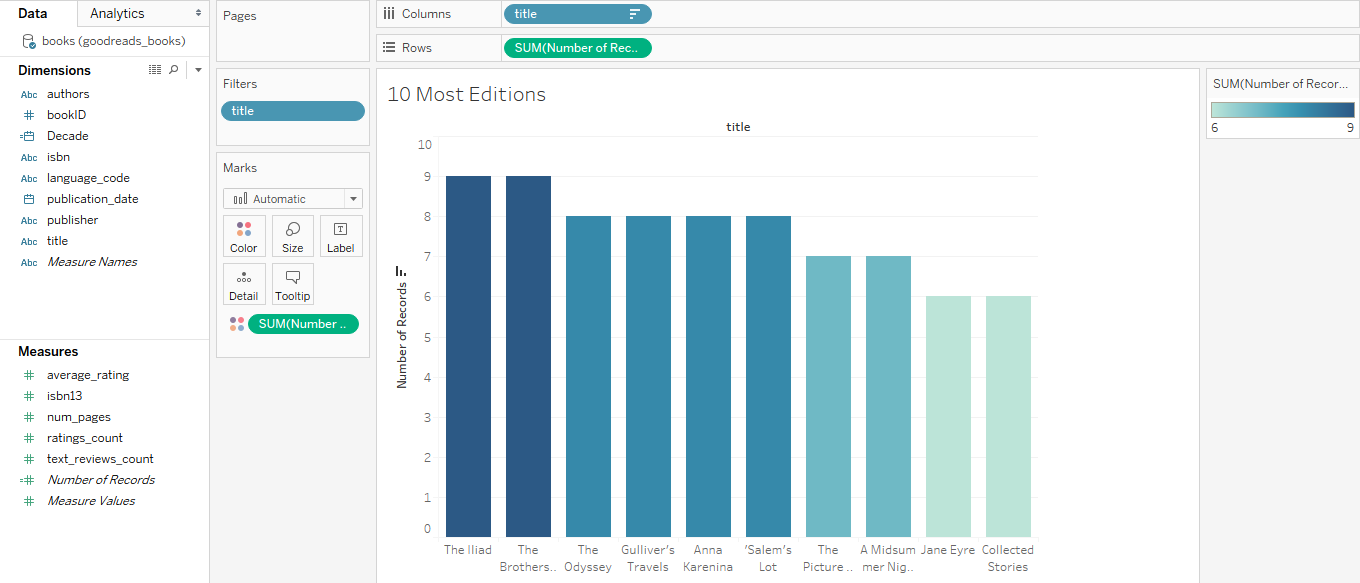
The distribution of languages can be shown a bit differently.



**Fig 22: Distribution of Languages Worksheet**

Here, knowing that only few of the categories in ‘language\_code’ occupies most of the data, it is plotted on a pie chart. Here, ‘language\_code’ is the converted dimension and sum of ‘Number of Records’ the measure. Some categories displayed as numerical values have been filtered out. Clearly, it can be noted that generic English is the most published language. Whereas English (US), Spanish and English (UK) follows respectively.

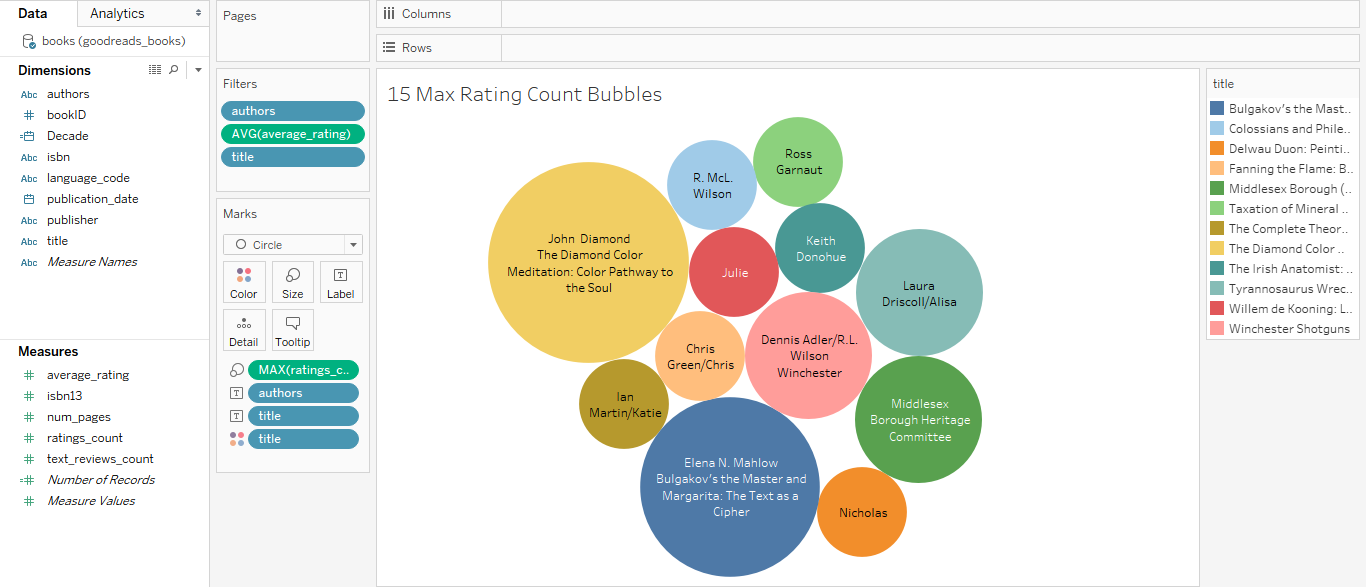
Next is finding out books having the greatest number of editions.



**Fig 23: 10 Most Editions Worksheet**

Here, ‘title’ is the dimension and sum of ‘Number of Records’ as the measure. As a default recommendation from Show Me, the data is plotted on a bar chart. Here, the sum of ‘Number of Records’ has been given a Blue-Teal color marker to observe the variance and ‘Number of Records’ has been sorted descending. Furthermore, the ‘title’ has been filtered to display only Top 10 ‘Number of Records’. The results show that ‘The Iliad’ and ‘The Brothers Karamazov’ has highest number of editions (9), followed by ‘The Odyssey’ with 8.

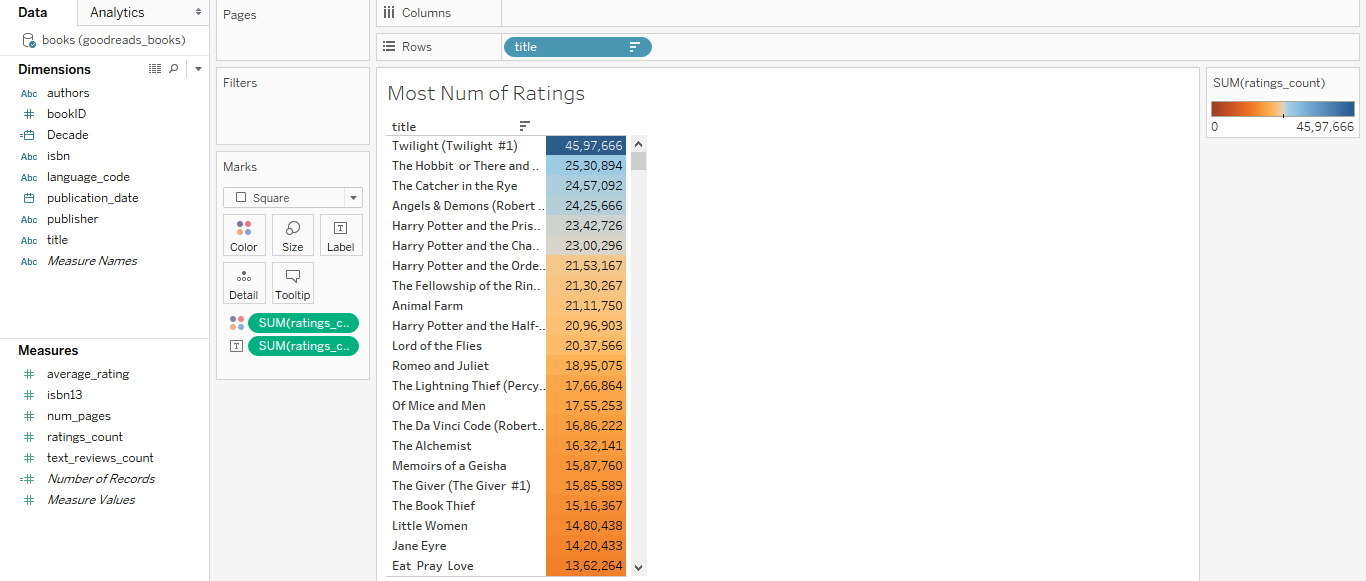
Next is to check the max rating counts in terms of average ratings.



**Fig 24: 15 Max Rating Count Bubbles Worksheet**

To accomplish that, the max of ‘ratings\_count’ as a measure has been plotted on a bubble chart with the book ‘title’ and their corresponding authors labelled on bubbles. Here, conditionally only 15 values with the highest ‘average\_rating’ has been shortlisted. Furthermore, some titles denoting as ‘NOT A BOOK’ have been filtered out. As seen on the chart, ‘The Diamond Color Meditation’ leads with max count as 5, followed by ‘Bulgakov’s Master and Margarita’ with 4.

At last, it would be an essential study to check which book is associated with the greatest number of ratings.



**Fig 25: Most Num of Ratings Worksheet**

Here, the book ‘title’ as dimension are simply described on text table with the sum of ‘ratings\_count’ as the measure. A color palette has been assigned to the measure and the table is sorted by values descending. It is quite evident that ‘Twilight’ leads followed by ‘The Hobbit’.

## 3.3. Questions

**Q1. Which year was the most successful in terms of book ratings for three authors with most books in the dataset?**

**Q2. What is the relationship of number of ratings, number of reviews and average rating with book length?**

**Q3. Which are the best publishers in terms of average ratings of books and within that the top 3 with maximum ratings and review counts? Which top 10 publishers publishes books with more than 1000 pages, on average?**

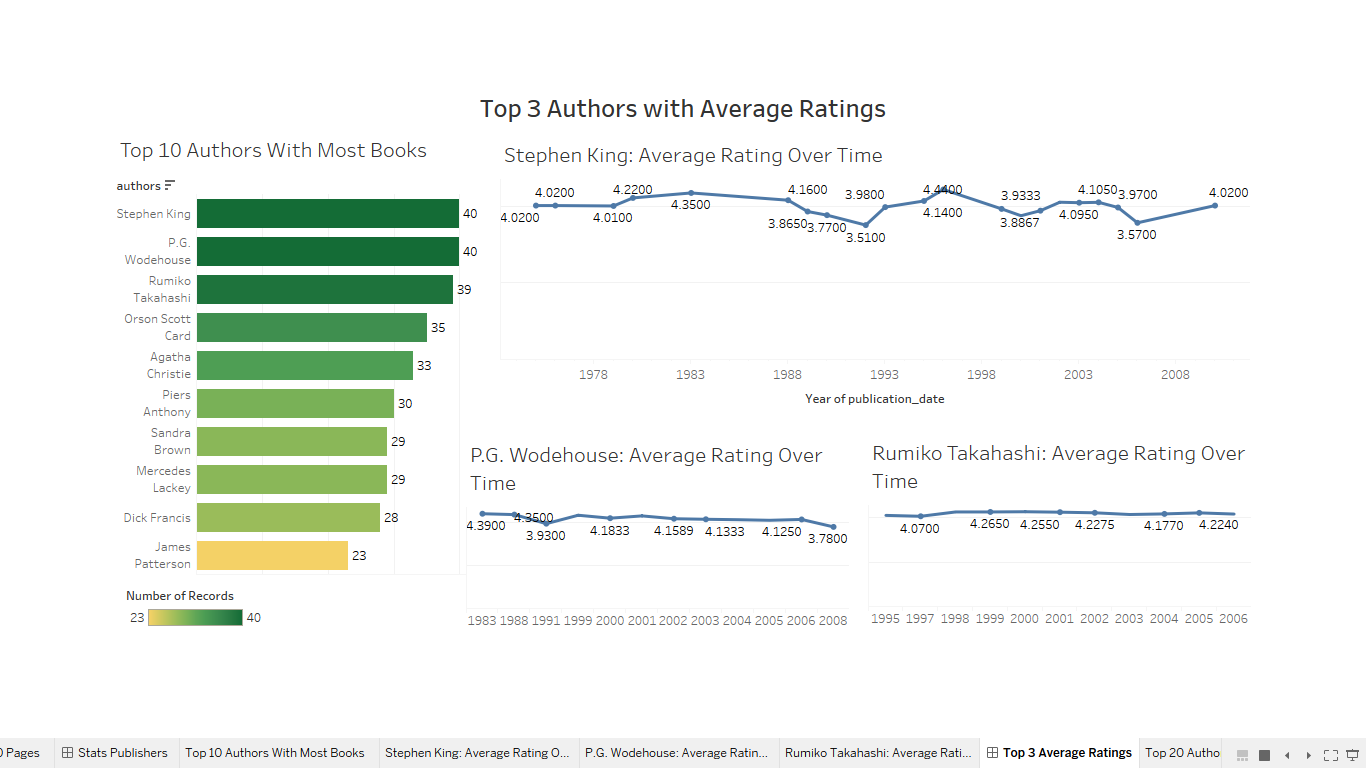
**Q4. Which authors, publishers and language received the greatest number of reviews? Furthermore, which was the best year for books to receive maximum of reviews, on average?**

**Q5. People are generally quite busy, so which books would be easier to finish reading?**

## 3.4. Visualized Answers

**Answer 1.**

To answer this question, foremostly, the ‘authors’ as the dimension needs to be plotted with the sum of ‘Number of Records’ plotted on a bar graph. Here, only top 3 authors are required. Then for each entry, a timeline of year in ‘publication\_date’ as a measure has to be plotted with average of ‘average\_rating’ as another measure plotted on a line chart.



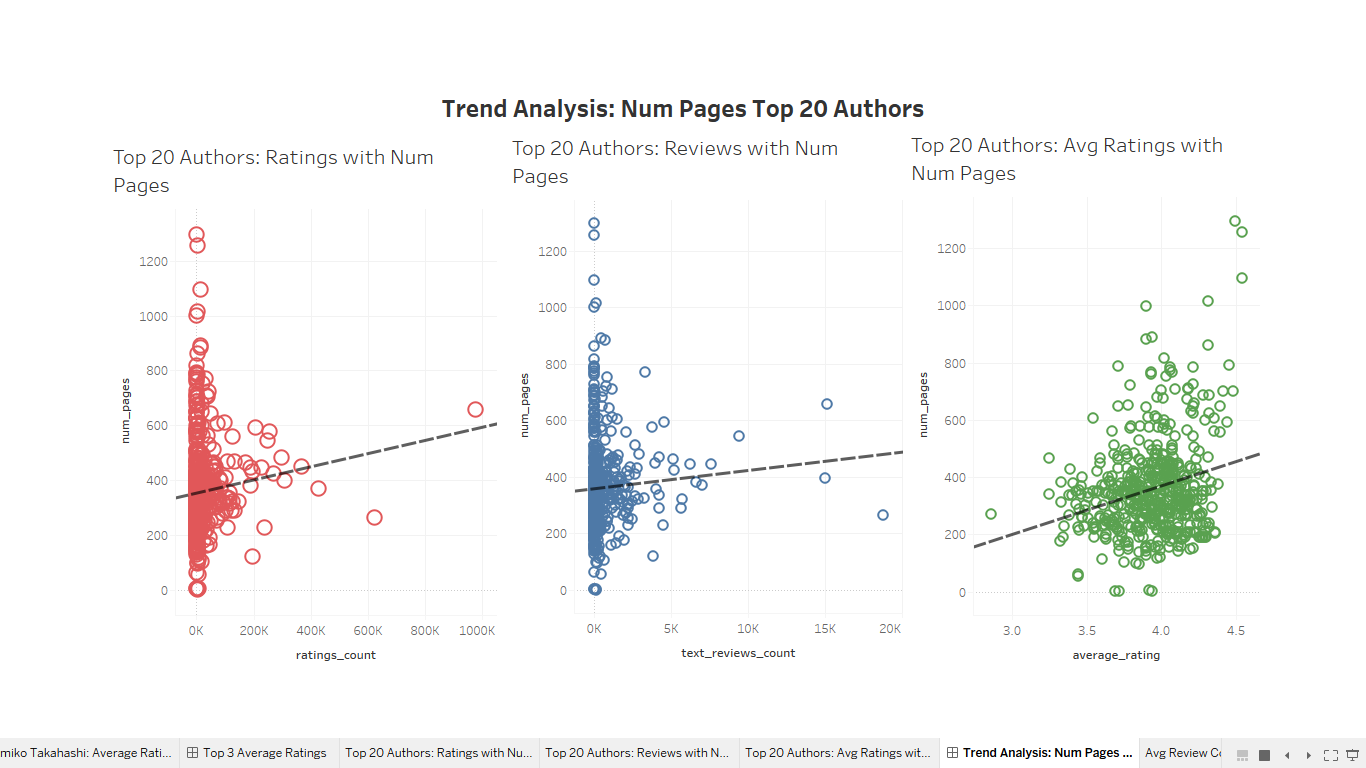
**Fig 26: Top 3 Authors with Average Ratings Dashboard Presentation**

The dashboard presentation clearly describes the Top 3 authors with most books as Stephen King, P.G. Wodehouse and Rumiko Takahashi. Noting the line graphs, the highest average rating for each can be given as follows:

1. Stephen King: 4.44
2. P.G. Wodehouse: 4.39
3. Rumiko Takahashi: 4.27

**Answer 2.**

To answer this, three scatter plots with ‘num\_pages’ have been visualized. To make the overall picture clearer this analysis is done by filtering out with top 20 authors with highest sum of ‘Number of Records’.



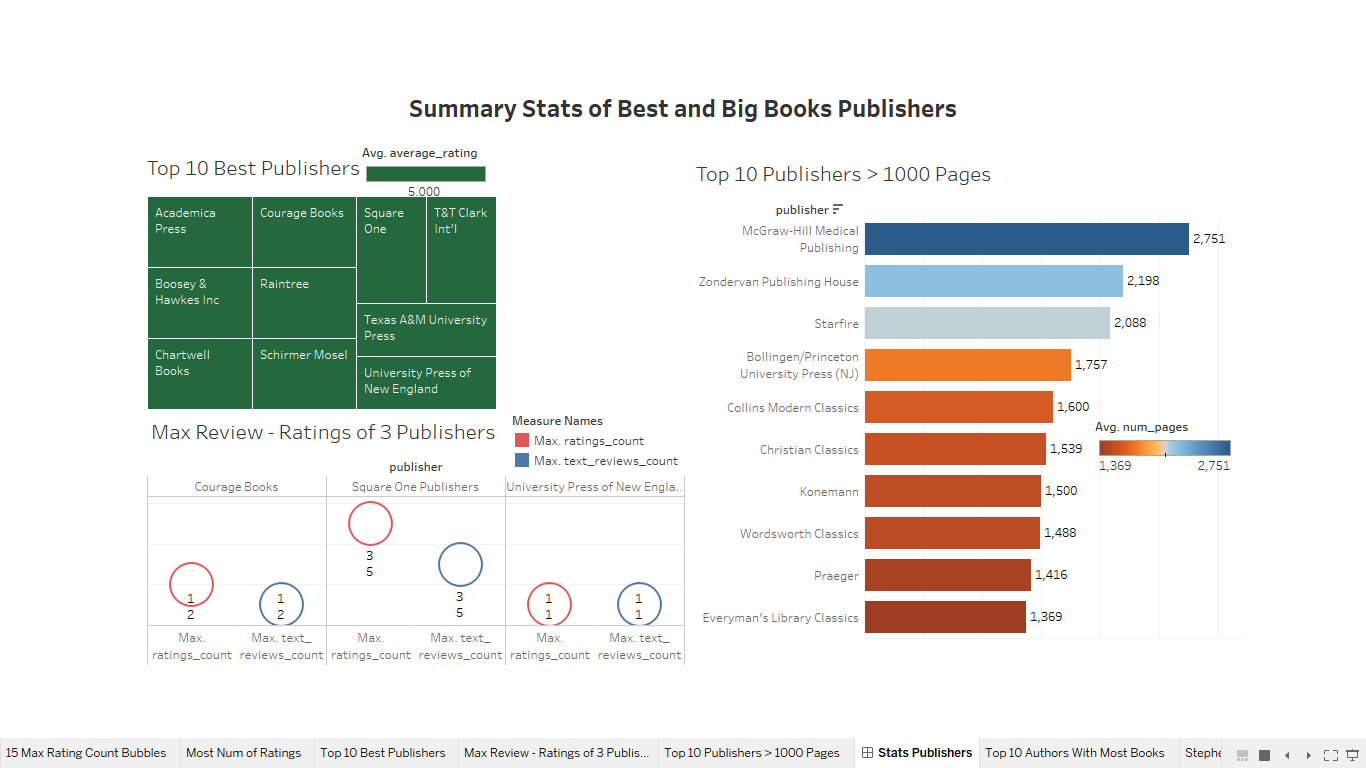
**Fig 27: Trend Analysis Dashboard Presentation**

Trend lines or lines of best fit have been added on top to answer effectively. The scatterplot for ‘ratings\_count’ with ‘num\_pages’ is highly skewed towards left. The line describes the trend to be increasing and hence there is a positive correlation between ‘ratings\_count’ and ‘num\_pages’. The scatterplot for ‘text\_reviews\_count’ with ‘num\_pages’ is also highly skewed towards left. The line describes the trend to be increasing but quite slightly and hence there is a little bit of positive correlation between ‘text\_reviews\_count’ and ‘num\_pages’. The scatterplot for ‘average\_rating’ and ‘num\_pages’ is distributed centrally. The line describes the trend to be increasing and hence there is an evident positive correlation between ‘average\_rating’ and ‘num\_pages’.

**Answer 3.**

To answer ‘publisher’ with > 1000 average of ‘num\_pages’, ‘publisher’ as a dimension is plotted with ‘num\_pages’ as measure on a bar plot and top 10 vales with highest average of ‘num\_pages’ are filtered.

To answer best publishers in terms of average ratings, a treemap is generated displaying top 10 ‘publishers’ with highest average of ‘average\_rating’. After that on a new sheet filtering out those top 10 publishers, side by side circle plots are made with maximum of ‘ratings\_count’ and maximum of ‘text\_review\_count’ as measures. Furthermore, the ‘text\_review\_count’ is filtered out of zero values.

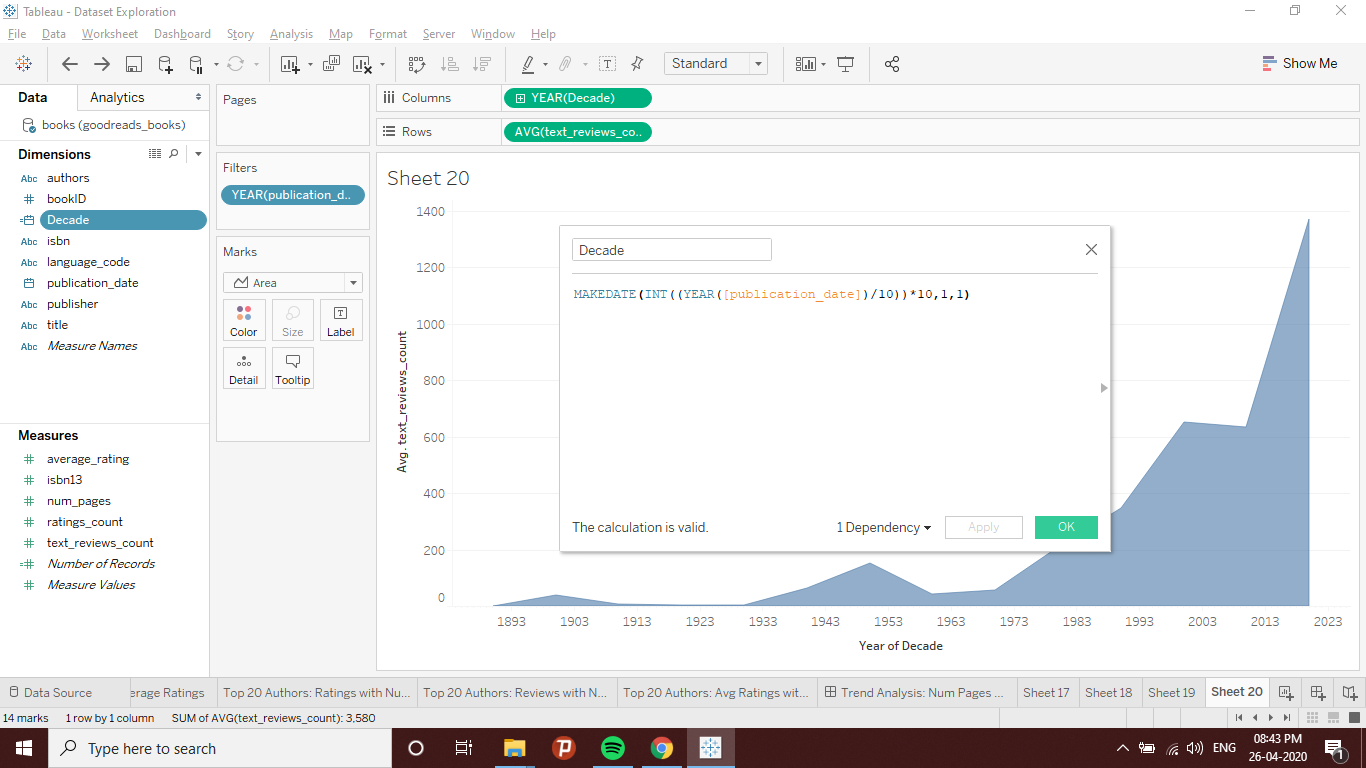


**Fig 28: Summary Stats of Best and Big Books Publishers Dashboard Presentation**

As shown in the dashboard presentation, McGraw-Hill Medical Publishing leads among the top 10 publishers with average ‘num\_pages’ greater than 1000. And Square One Publishers leads for both maximum ratings and review counts.

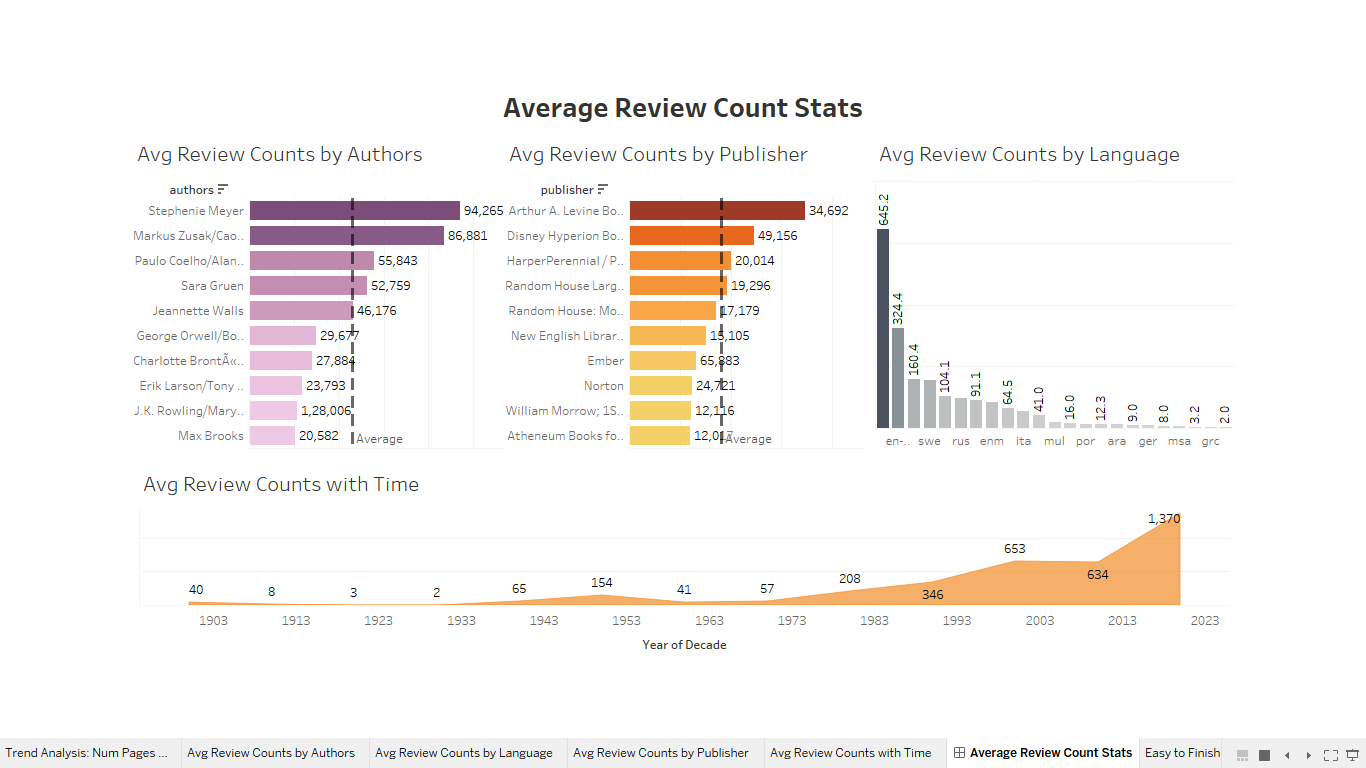
**Answer 4.**

To answer this question, three barplots with average of ‘text\_reviews\_count’ as the measure are plotted with ‘authors’, ‘language\_code’ and ‘publisher’ as dimensions respectively. Plotting ‘publication\_date’ would not be very feasible for exploratory purpose as it contains a large range of date from 1890 to 2020. So instead, a new calculated column called ‘decade’ is formulated as follows:



**Fig 29: Calculated Field**

Then ‘decade’ as in year as a measure is plotted with average of ‘text\_reviews\_count’ as another measure on a shaded line chart.

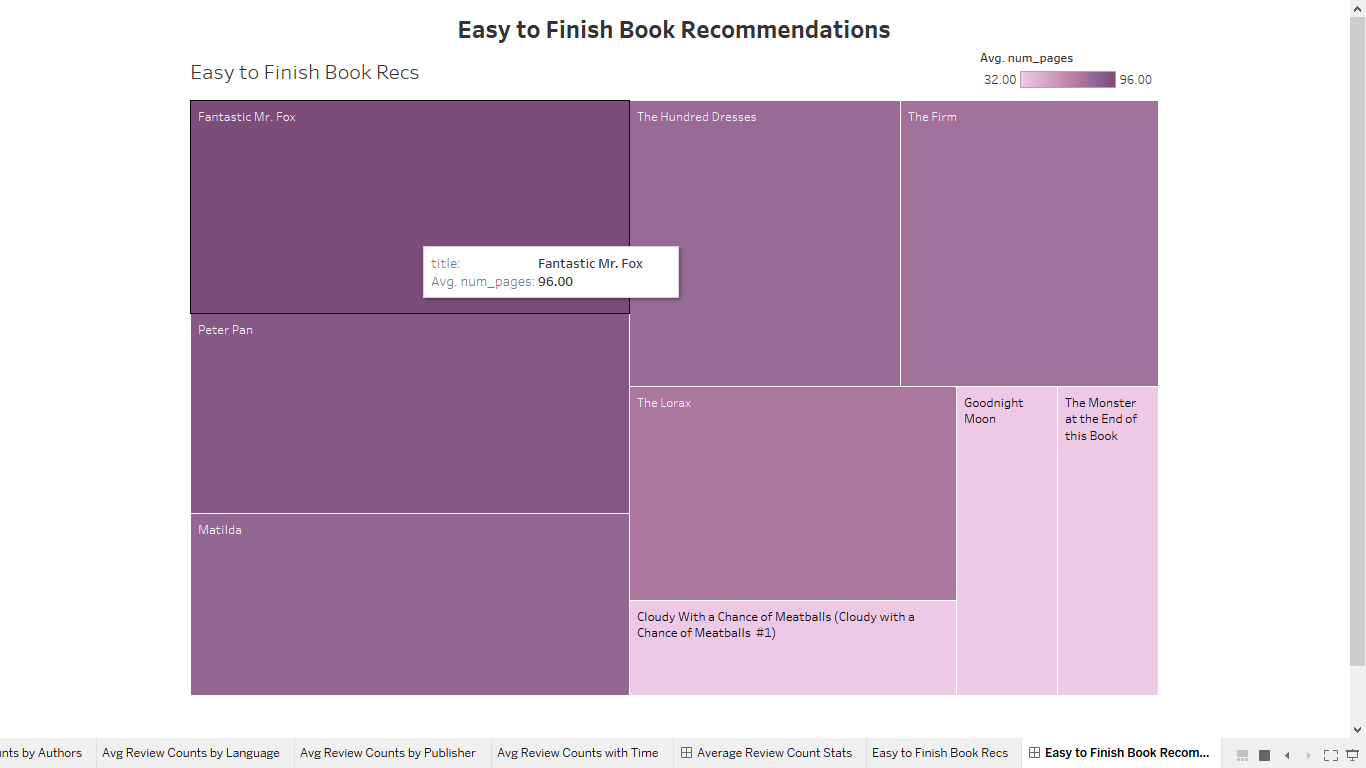


**Fig 30: Average Review Count Stats Dashboard Presentation**

As shown on the dashboard presentation, Stephenie Meyer leads among authors, Arthur A. Levine Books among publishers and English among languages. Also note that books published in range of 2010 - 2020 received highest number of reviews, on an average.

**Answer 5.**

To answer this last question, pertaining to all the dimensions and measures in the dataset, it is most logical to recommend a book that is shorter to finish it easily. So here, a treemap is plotted on average of ‘num\_pages’ as the measure. Moreover, for the better experience of the reader, it is filtered on ‘average\_rating’ of atleast 4 and ‘text\_reviews\_count’ of atleast 2000 hits. The results are as follows:



**Fig 30: Easy to Finish Book Recommendations Dashboard Presentation**

# REFERENCES

*Goodreads-books | Kaggle* (no date). Available at: https://www.kaggle.com/jealousleopard/goodreadsbooks (Accessed: 26 April 2020).

*Indicators | Data* (no date). Available at: https://data.worldbank.org/indicator?tab=all (Accessed: 23 April 2020).

*Tableau Dashboard Tips [Top 10 Tableau Dashboard Design Tips] - YouTube* (no date). Available at: https://www.youtube.com/watch?v=Lu0jrymqOGM (Accessed: 25 April 2020).

*Tableau Tutorial - Tutorialspoint* (no date). Available at: https://www.tutorialspoint.com/tableau/index.htm (Accessed: 25 April 2020).