**Multi-Dimensional Visualization of Student Book Choice**

Dhanya Mathew1,2, Khushboo Khandelwal3,4, Nick Stames1,2,Supriya Srinivasa3,4, Swarnamouli Majumdar3,4

1 Indiana University Bloomington, IN 47408, USA

2School of Informatics, Computing and Engineering, Indiana University Bloomington, IN 47408, USA

3Indiana University-Purdue University Indianapolis, IN 46202, USA

2School of Informatics and Computing, Indiana University-Purdue University Indianapolis, IN 46202, USA

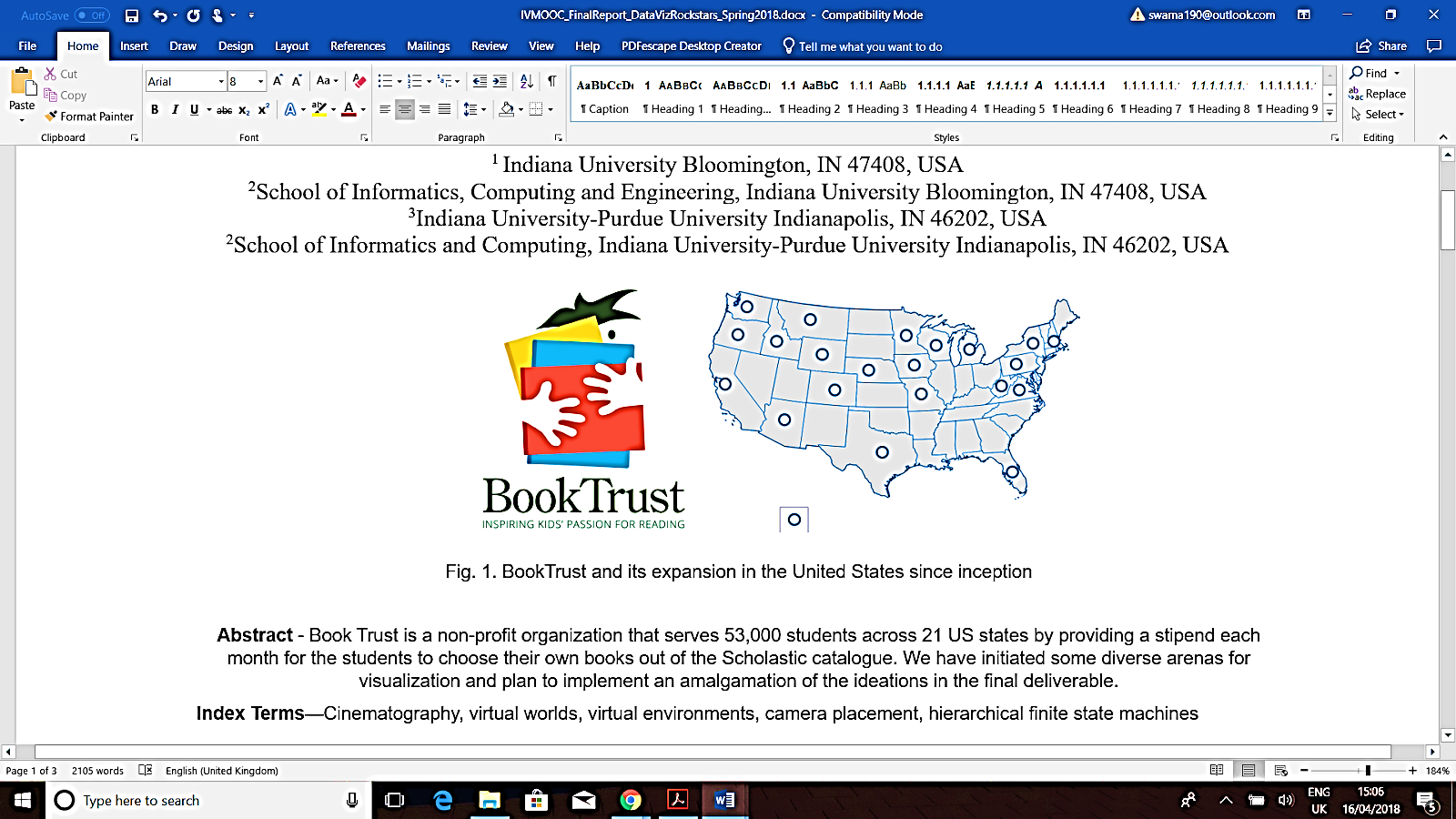


Fig. 1. Book Trust expansion in the United States since inception

**Abstract** - ***Book Trust, the Client (as referred in this document), is a non-profit organization that serves 53,000 students across 21 US states by providing a stipend each month for the students to choose their own books out of the Scholastic catalogue. Our research initiated some diverse arenas for visualization and implemented an amalgamation of the ideations through visualizations as final deliverables. We collected dataset(s) from Client, and processed data for preparation, and transformation; this was followed by data analysis, including dataset(s) summary and descriptive statistics on key business metrics of Client business model. The data analysis insights were leveraged to build multi-dimensional dynamic visualizations on Tableau, including location analysis, pareto analysis, value analysis, word cloud comprising Client served states, and etc.,. We also designed a temporal bar graph entailing top book titles using Sci2 open source tool. The product was shared with the Client, for deployment within their existing Tableau environment.***

***Index Terms*** *– book titles; data analysis; data pre-processing; descriptive statistics; dataset summary; data transformation; exploratory analysis; dynamic visualizations; geolocation visuals; value analysis; pareto analysis; word cloud; non-profit; temporal bar graph; Si2, Tableau*

**Introduction**

This project demonstrates the implementation of a multi-dimensional visualization dashboard on Tableau software, which narrates a story commencing with the inception of Client, its growth, geolocation and value analysis, statistical analysis of data, pareto analysis and other significant findings. We intended to create a value proposition, which Client can leverage post deployment as a starting point to build high quality visual dashboards, enabling them to construct a future sales model based on our findings and visualizations

The sections, which follow entail detailed information on project exposition, visualization goals, envisioned visualizations, data analysis, and completed final visualizations.

**Authors**

***Dhanya Mathew*** *is a graduate of* ***MS Data Science*** *at Indiana University, Bloomington, Indiana. Email: dmathew@iu.edu*

***Khushboo Khandelwal*** *is a graduate candidate of* ***MS Applied Data Science*** *at Indiana University-Purdue University Indianapolis, Indianapolis, Indiana*

*Email: kkhandel@iu.edu*

***Nick Stames*** *is a graduate candidate of* ***MS Data Science*** *at Indiana University, Bloomington, Indiana. Email: nstames@iu.edu*

***Supriya Srinivasa*** *is a graduate candidate of* ***MS Applied Data Science*** *at Indiana University-Purdue University Indianapolis, Indianapolis, Indiana*

*Email: supsrini@iu.edu*

***Swarnamouli Majumdar*** *is a graduate candidate of* ***MS Applied Data Science*** *at Indiana University-Purdue University Indianapolis, Indianapolis, Indiana*

*Email:* *swmajumd@iu.edu*

# Exposition

## About BookTrust (Client)

Client is a non-profit organization that serves 53,000 students across 21 US states by providing a stipend each month for the students to choose their own books out of the Scholastic catalogue.

## Dataset(s) Explanation

Every month BookTrust receives data from Scholastic, their business partner who supplies book. The data essentially includes information on books order transactions, including details about order number, book titles, book price, book quantity, Client order value, Total order value, over stipend order values, catalogue identification, book ISBN number, states, districts, schools, teachers, and etc., For this project, Client has replaced the critical business data associated with the districts, schools, and teachers name with numeric codes also known as key party identifiers to maintain the privacy and confidentiality of the involved parties. The data that we received from the Client, include orders from September 2017 until February 2018.

In total, we received three (3) excel files, which has data and metadata associated details, including records. These files are also a database table in itself. We also received a data dictionary file in word document to read each data file. The list of the files is summarized below:

**File: DSet02.xls(Party Table)**; This file contains essential information about party table also used in measuring KPIs. The party table includes details about parties where & with whom Client is having its modus operandi; list of states, list of school districts, list of schools within the school districts, and list of teachers.

**File: DSet01.xls(Order Table)**; This file has detailed information on each order summary; order date, order placed by schools, order placed by school teacher, order total, book trust order total, over stipend order total, and etc., These orders data are for the School year 2017-2018, starting from September 2017 until February 2018.

**File: OI2.xls(Order Items Table);** This file has detailed information on line items for each order transaction, including book titles, book ISBN, book quantity, book unit price, total book unit price, catalogue code, mapped to order id of books.

**File: Data Notes.doc(Data Dictionary)**; This file comprises exhaustive information on how to interpret the three data files also knows as tables, essentially entailing information about table relationships, metadata definitions and relationships between tables.

## Visualization Goal

Our goal for this project is to provide our Client with detailed analysis, including visualizations leveraging dataset(s). Our visualizations will provide a closer look into their data that can be used to extract meaningful insights.

Summarized below key objectives of our project, which we have accomplished:

* Performed data analysis to gather preliminary insights on trends and patterns within the dataset(s) to share key insights associated with orders, and books from September 2017 through October 2018
* Developed a multi-dimensional visualizations product on Tableau that will support the above-mentioned objective for future usage

# Background

Our Client has indicated that they currently have Tableau dashboards deployed in their environment. Some of them are enlisted below.

1. Key indicators by month (Number of States, Schools, Teachers, Students, Order Totals) - Numeric table
2. Orders distribution by day of month - Line graph
3. Student count by teacher by month - Numeric table
4. Total order amount (in Dollars [$]) over time - Line chart
5. Top 10 ordered titles filtered by book price - series of Numeric tables.

Our project aimed to complement their existing dashboards, as well as implement additional data analysis and visualizations as described in the Visualization goal section.

# Data Pre-processing

Post finalization of dataset(s), it was processed for table normalization and data transformation. The relationships we interpreted among the data variables are as follows:

1. *Teacher* - Instructor in charge of a classroom.  Orders are tied to a teacher.
2. *School* - Teachers teach at one or more schools.
3. *District* - Schools belong to one school district.
4. *State* - Districts exist inside a state.
5. *Stipend* - BookTrust pays for a student’s books up to this amount per order

The data pre-processing steps included these steps: party table was normalized to organize details on unique states, unique districts, unique schools, unique teachers via party id field within the party table. Order table was mapped with states, districts using school id as foreign key to party table. Order items table was mapped with states, districts, schools using order id field as foreign key to order table.

# Data Analysis

## Dataset(s) Exploration and Summary

* + 1. **Party Table: Client Master Data on Party Type since Inception of BookTrust**

Total Number of Parties: 5 (Party 1 is Teacher; Party 2 is School; Party 4 is Districts; Party 5 is Book Trust; Party 6 is Scholastic); Number of States: 21; Number Districts: 87; Number of Schools within Districts: 244; Number of Teachers: 7,295

* + 1. **Order Table: Orders details from September 2017- February 2018**

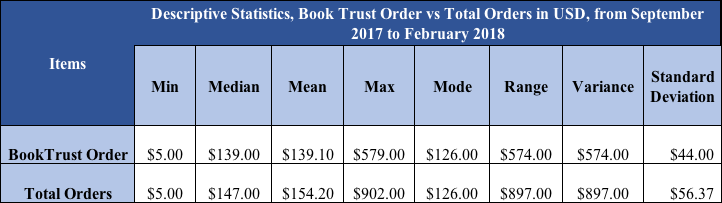
Total Number of Orders/Rows: 14,286; Number of States: 21; Number of Districts: 61; Number of Schools within Districts: 174; Number of Teachers: 2,608

* + 1. **Order Items Table: Each Order Items detail from September 2017-February 2018**

Total Number of Rows/Orders Line Items: 577,652; Number of States: 21; Number of Districts: 61; Number of Schools within Districts: 174; Number of Teachers: 2,608; Number of Unique Book Titles: 4,092 with 4,047 ISBNs; Number of Total Books: 585,256

## Descriptive Statistics

## Book Trust Order Total and Total Orders Analysis



*Fig 2. Comparative Analysis of Book Price Order& Total Order, from September 2017 to February 2018*

Fig. 2., is a comparative analysis of the Book Trust Order Total, and the Total Orders, from September 2017 to February 2018. We computed exploratory analysis of the Total orders and the Book Trust orders from September 2017 to February 2018, which displayed some key insights from the data. The maximum Book Trust Order Total was $579 while the maximum total order was $902, with an enhanced margin of $323. The mean order value of both BookTrust Order, and Total Orders stand at $139 and $154 respectively, with minimal standard error and similar mode value of $126.

## Book Price and Book Quantity Analysis



*Fig 3. Scatter Plot to display Unit Book Price and Order Quantity, from September 2017 to October 2018*

Fig 3., Is a scatter plot displaying the relationship between unit book price and order quantity. The correlation between these variables are somewhat negative. The ‘X’ axis represents Unit Book Price and ‘Y’ axis represents Order Quantity. Interestingly, the relationship is a weakly correlated as the coefficient value is somewhat around (0.116).



Fig. 4. Descriptive statistics summary on Book Price (USD)

Total 6,720 books ordered for unit book price of $0, and total 3 books ordered for unit book price of $100.

# Data Visualization

After data pre-processing, and data analysis, our team commenced with implementing our envisioned ideations and sketches on Tableau and produced the following dynamic visualizations. The visualizations below have been developed on Tableau software and screenshots of the visualizations have been provided below.

**5.1 Location Analysis**

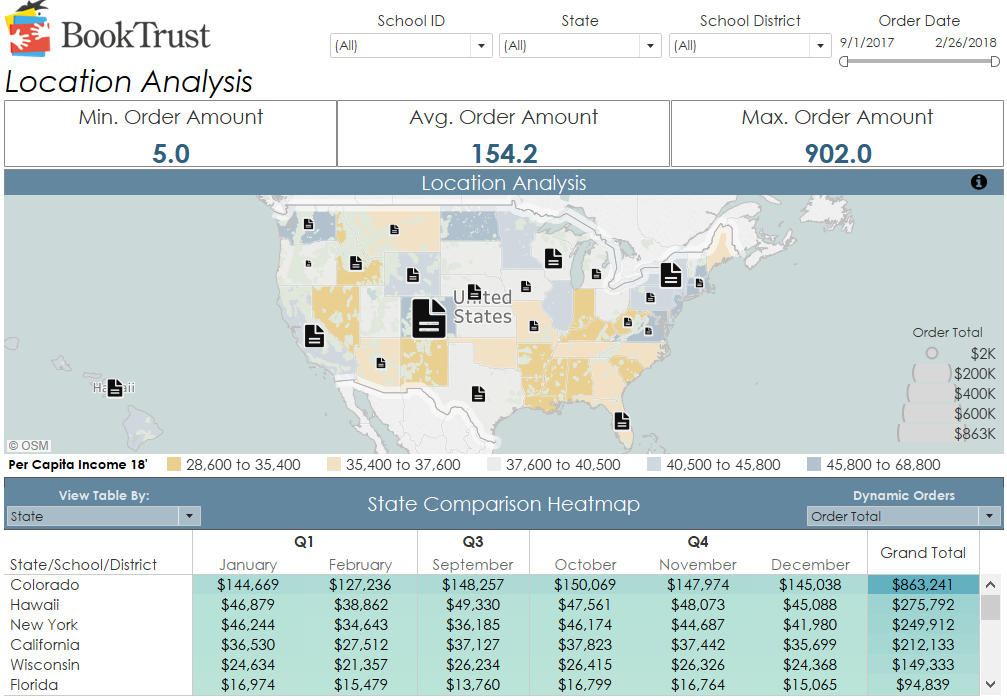


Fig. 5. Location Analysis of orders in combination of geolocation, and heatmap

**Visualization Description:**

The purpose of this dashboard is to provide the end user with location-based analytics. Considering order totals, we have added the 2018 Income per capita base layer to our map to provide immediate insights into state level demographics.

The heatmap based approach is used to show orders by location. The Lighter colors represent lower orders totalled by state, whereas the darker colors represent higher order totals. The bottom table (heatmap) allows the user to get an in-depth view of the order totals with a Quarter over Quarter view. A parameter control is added which allows the end user to view the total by State/District/ School.

The Key Metrics used in this visualization areState Analysis, School Analysis and District Analysis. The Key Filters are School ID, State, District and the Order date, the order date allowing custom values. The final Measures we intend to output are the *BookTrust Order total* and the *Total orders*.

**5.2 Pareto Analysis**

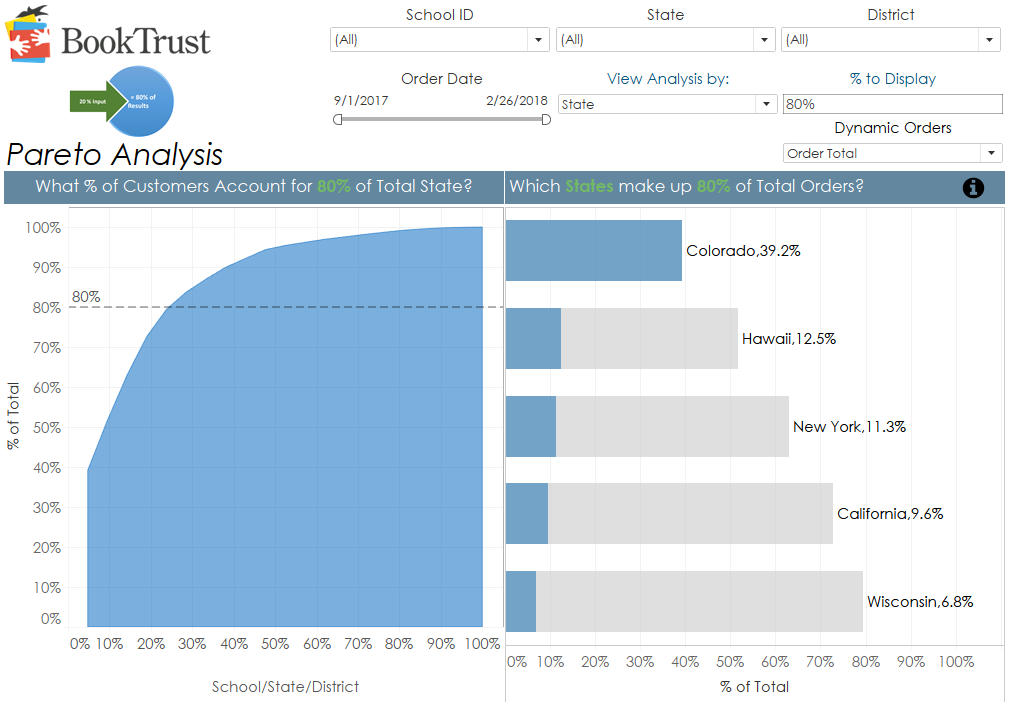
****

Fig. 6. Pareto Analysis based on the ‘*Pareto Principle*’

## Visualization Description:

The Pareto Analysis provides a visual view into the concept of the Pareto Principle. In short, the Pareto principle, more popularly, is known as the ‘80/20 rule’. It is the *law of the vital few*. Which interprets into the fact that implementation 20% of the input yields 80% of the total results. In this case we have allowed the end user to dictate the analysis to show School/State/District as the *driving analysis.*

The Key Metrics used in this visualization areState Analysis, School Analysis and District Analysis. The Key Filters used in this visualization are School ID, State, District, Order date, which allows for custom range and View Analysis, which allows analysis by School/ State/District. The final Measures we intend to output are the *BookTrust Order total* and the *Total orders*.

**5.3 Value Analysis**

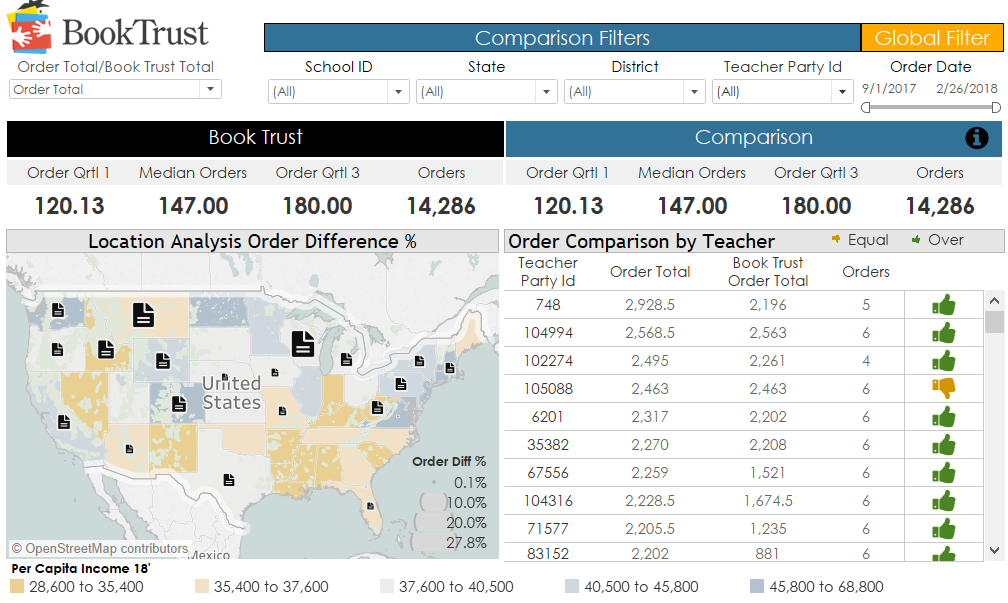


Fig. 7. Value Analysis based on teacher-related metrics

**Visualization Description:**

The Value analysis visualization provides a representation on teacher related metrics. The top layer of the dashboard allows our Client to compare state, school, and district to the rest of the organization (i.e., Book Trust’s) practice.

In the map view, it provides insight into state level percentage difference demonstrating which states ordered more books with the book trust stipend and which used only the allotted amounts.

The key metrics of this visualization are the ‘First Quartile’, the ‘Median’ and the ‘Third Quartile’. The key filters are school id, state id, district it, and the order date, the order date allowing custom values. The final output measures as per our goal are Book Trust order total, order total, percentage difference [the difference of the order total and the Book trust total], *equal stipend* order which are indicated by a *downward thumb* and shows no difference between order total and the BookTrust order total, thus displaying no overage in order amounts over the allotted stipend amount, and *over stipend* orders**,** which areindicated by an *upward thumb* displaying a positive increase over the BookTrust order total, thus concluding to an overage and teachers ordering more books than just the allotted stipend amount.

**5.4 Word Cloud**



Fig. 8. State Cloud of Top Performing US States by Book Orders

**Visualization Description:**

The state cloud dashboard provides the end user with quick analysis of the top performing states based on order quantity and the number of unique books ordered. The key metrics of the state cloud visualization are the order quantity and the number of unique books ordered by the states.

**5.5 Temporal Bar Graph using Sci2 tool**

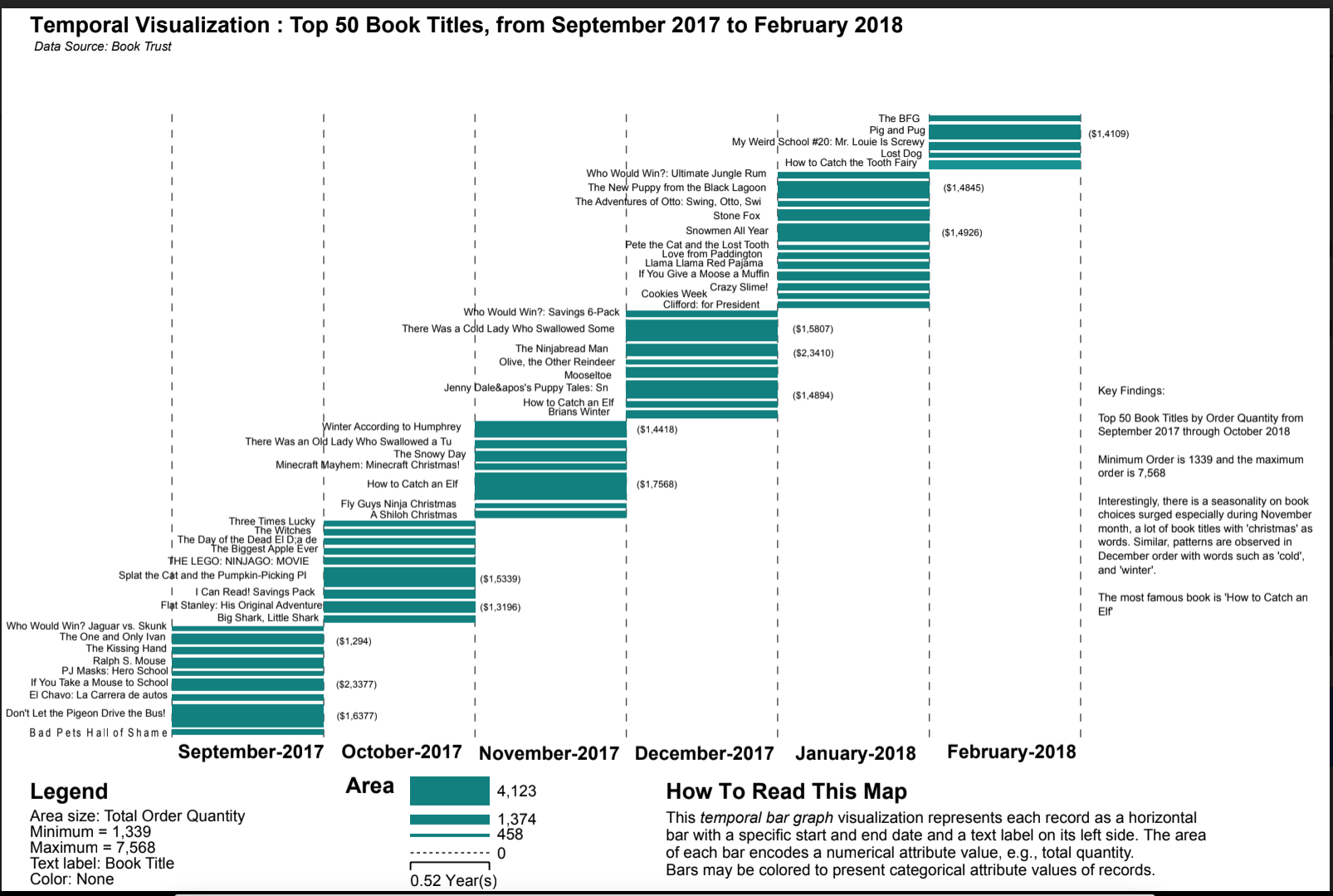


Fig.9. Temporal Bar Graph by Monthly Book Orders for School Year 2017-2018

**Visualization Description:**

Fig.9. The temporal bar graph depicts, the top 50 book titles by order quantity, from September 2017 through October 2018. The bottom 50th book **‘Fly Guys Ninja Christmas’** 1,339 copies were ordered, and the top 1st book **‘How to Catch an Elf’** 7,568 copies were ordered.

Interestingly, there is a seasonality that can be interpreted after this graph, especially during November, a lot of book titles with 'Christmas' as words have been on demand, and similar patterns were observed in December month order with words such as 'Cold', and 'Winter'.

# Key findings

* 1. Overall, it can be concluded that if unit book price is more than less number of books are ordered
  2. Top 5 books are: ***How to Catch an Elf, Don't Let the Pigeon Drive the Bus! There Was a Cold Lady Who Swallowed Some, Splat the Cat and the Pumpkin-Picking Pl, Snowmen All Year***
  3. Colorado is the top state in terms of order quantity as well as order price. New York is in second place while considering top 10 books per state w.r.t book price. But when we choose top 50 books per state, Hawaii comes to second place may be because New York has more low-priced books in their top 50.

# Conclusion

Our project report finding lead us into concluding that few of the States have scored very well on various metrics like book orders, over stipend book orders and teacher related metrics. An important of the Pareto Analysis is that the states of Colorado, Hawaii, New York, California and Wisconsin make up 80% of the total book orders. The temporal bar graph is quite unique in its demonstration of the seasonality of book title orders spread across months. Therefore, the measure of book orders by various metrics have been captured from a multi-dimensional perspective throughout the course of this research project.

# Acknowledgments

*This research cum project was initiated with the whole-hearted support of Professor Emeritus Katy Borner of Indiana University Bloomington and was conducted as part of our project work for IVMOOC. We would also like to thank our Client BookTrust for supporting us all through the project execution and sharing their invaluable feedback and suggestions. We would also like to extend our gratitude to the Peer Feedback session from IVMOOC for sharing their wisdom with us during this research. We are immensely grateful to Professor Katy and Michael Ginda for their comments on an earlier version of the manuscript. We would like to take the opportunity to thank Cyberinfrastructure for Network Science Center for the Sci2 Software we used along with Tableau.*

# References

1. Börner, K., & Polley, D.E., 2014. *Visual insights: a practical guide to making sense of data.* Cambridge, MA: The MIT Press.
2. Börner, K. 2015. *Atlas of science: visualizing what we know.* Cambridge, MA: The MIT Press.
3. Bingham, G. & Hall-Kenyon, K. (2013). Examining teachers’ beliefs about and implementation of a balanced literacy framework. Journal of Research in Reading, 36(1), 14-28.
4. Gambrell, L. B. (2014). Seven rules of engagement: what’s most important to know about motivation to read. Reading Teacher, 65(3), 172–178.
5. Katz, I., & Assor, A. (2007). *When choice motivates and when it does not.* Educational Psychology Review, 19(4), 429–442.