Data Governance with Analytics

[1]Ameya Parkar, [2]Bhumika Dalal
[1]Assistant Professor, V. E. S Institute of Technology, Mumbai, India
[2]Student, V. E. S Institute of Technology, Mumbai, India

Abstract: ***"Data governance is an emerging trend in enterprise information management" [4]. The purpose of data governance is to minimize the cost and risk and also to increase the value of data. It requires data to be correct and also ensures the quality of data. Organizations today struggle to organize and utilize the data, assure its quality, and turn it into measurable business value. The purpose of this research paper is to analyze the data and bring useful insights that will help in better decision-making, enhanced operational efficiency, and increased revenue for data governance. For better decision making and to ensure that the data is fit for purpose, the approach used in this paper is Exploratory Data Analysis.***

Key Words **– Data Governance, Exploratory Data Analysis, Analytics, Data mining**

I. INTRODUCTION

Data governance is a set of management behaviors about data usage in an organization [2]. Data governance is control over the management of data and helps in maximizing the value of the organization's data asset and minimizing the risks related to data. As the volume of data is increasing from diverse sources there needs to be a check on data before the decisions are being made based on incorrect and inconsistent data. Data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data [6]. It is like a framework that helps the organization in the management of data which helps in cross-functional team collaboration. Governance is not a one-time project, but a constant program. Data governance focuses on data as a strategic enterprise asset [5]. It is important to get insights about data, how are different fields correlated to each other before making decisions based on the available data. It will also help to ensure that the data is fit for purpose and improve the decision-making process. These insights about data can be brought using Exploratory Data Analysis (EDA). New tools for raw data characterization of these datasets through EDA are required to suggest initial hypotheses for testing [3].

II. PROBLEMS AND CHALLENGES

Organizations implementing analytics as an integral part of their decision making typically have several stakeholders: managers, analytics practitioners, and data management practitioners [1]. Stakeholders face major problems which do not lead to effective decision making. From a data management perspective it becomes difficult to determine how the data is correlated when there are incomplete or inconsistent data. Due to limitations in the infrastructure of data governance the insights of data and its impact can't be provided to the stakeholders. Without analytics it would be difficult for Data Stewards to find the Critical Data Elements (CDE), so by using analytics it becomes easy to find the CDE's in the data. Data management practitioners are struggling with numerous ad-hoc data requests from analytics practitioners and managers [1]. As there are limitations in time and cost it becomes a long process to determine an effective decision-making process. The problems and challenges can be overcome by using analytics within data governance.

III. DATA GOVERNANCE DEFINITION

Data governance specifies decision rights and accountabilities for an organization's decision-making about its data. Furthermore, data governance formalizes data policies, standards, and procedures and monitors compliance [5]. It helps to ensure the availability, integrity, usability, and security of data based on the standards and policies of the organization and provides a cross-functional framework for managing data as a strategic enterprise asset. The goals of data governance include reducing the cost of managing data, resolving issues within a data set, positioning data as a high-value asset as well as maintaining data policies and procedures, highlighting outliers in the data that deviate from corporate standards and provide feedback to data stewards. It is the core component of data management process and ensure that the data within the organization is consistent and is not misused. Data governance should include analytics to optimize the decision-making process. By providing more insights on data, effective decisions can be made. It enforces common definitions and standard formats boosting data consistency for business and helps to break the silos in an organization that has a separate transaction processing system and have a decentralized coordination. Data governance overcomes decentralized coordination by harmonizing data through a collaborative process with stakeholders from different units. Another benefit that data governance provides is the data quality along various dimensions such as completeness, conformity, uniqueness, timeliness, integrity, accuracy, and consistency.

The newly issued "Information Technology Service Governance" national standards set the information technology governance objectives in four aspects, namely, strategic consistency, risk control, operational compliance, and performance improvement [2]. From these objectives, we can see that the essential goal of governance is value and risk, and governance is the framework of decision-making and responsibility to encourage expected behavior [2]. There are various critical success factors of data governance which are as follows:

- Cross-functional involvement: It should involve the participation of stakeholders from different units.
- Alignment with business objective: Alignment of data and process with business objectives.
- Metrics: Data quality metrics help to measure the success of data governance.
- Policies and standards: Must ensure that the data is fit for purpose.
- Compliance monitoring: Assessed periodically to ensure that the procedures are followed.

IV. COMBINING ANALYTICS AND DATA GOVERNANCE

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods also EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. [7]. It performs an investigation on the data to discover patterns, spot anomalies, outliers, test hypothesis, and check assumptions with the help of graphical representation and summary statistics. It postpones the assumption on what kind of model the data follows rather a direct approach for the data itself to reveal its structure and model. EDA helps in finding what are we looking for, how we look at the data, and how we interpret the data. This graphical visualization helps in understanding the structure of data and reveal new and unsuspected insights. Reasons to use EDA:

• Helps to detect mistakes

• Determine relationships among the variables

• Preliminary selection of model

• Checking for assumptions

Organizations today struggle to properly organize and utilize data, assure its quality, and turn it into measurable business value. Effective data governance combined with analytics brings value within a short span of time. This helps to focus on establishing a strong master and transactional data governance organization that is responsive and adaptive. With the help of EDA it is possible to discover the data trends and make effective decisions using these insights. The dataset used here is BigMart Sales data and have collected 2013 sales data for 1559 products across 10 stores in different cities. The raw data has been plotted using different EDA graphs such as box plot, histogram, scatter plot, heatmaps, etc.

Figure 1(a) and 1(b) is an example of Univariate analysis. It is one of the simplest forms of analysis, where the data being analyzed includes a single variable. The purpose of univariate analysis is to find patterns that exist within it which describes the data. Figure 1(a) shows the distribution plot for the sales variable and the trends in sales. It shows that the distribution of sales has a positive skew and is right-skewed, right-tailed. Figure 1(b) shows the probability graph for sales compared against the theoretical set.

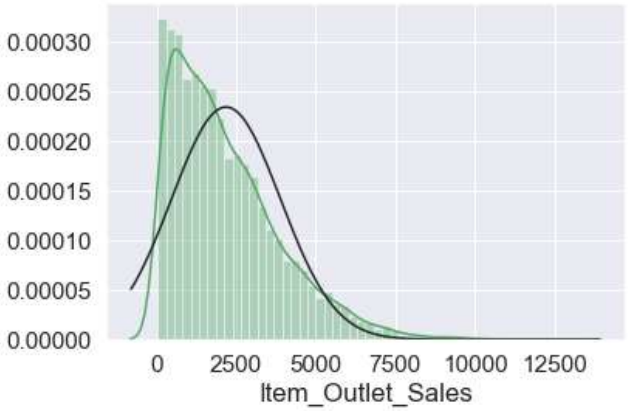

Figure 1(a). Distribution plot of Sales

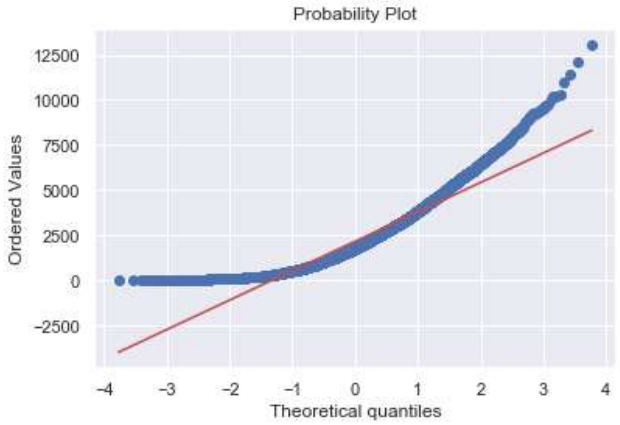

Figure 1(b). Probability plot of sales

Another example of univariate analysis is the histogram. It helps to discover the frequency distribution of continuous data whether the distribution is a normal distribution, skewed or there are outliers in the data. Figure 2 shows the frequency distribution of four variables in the data set. Item_MRP is a multimodal histogram distribution. Item_Visibility, Item_Weight, and Item_Outlet_Sales show a positive distribution and it is a right-skewed distribution.

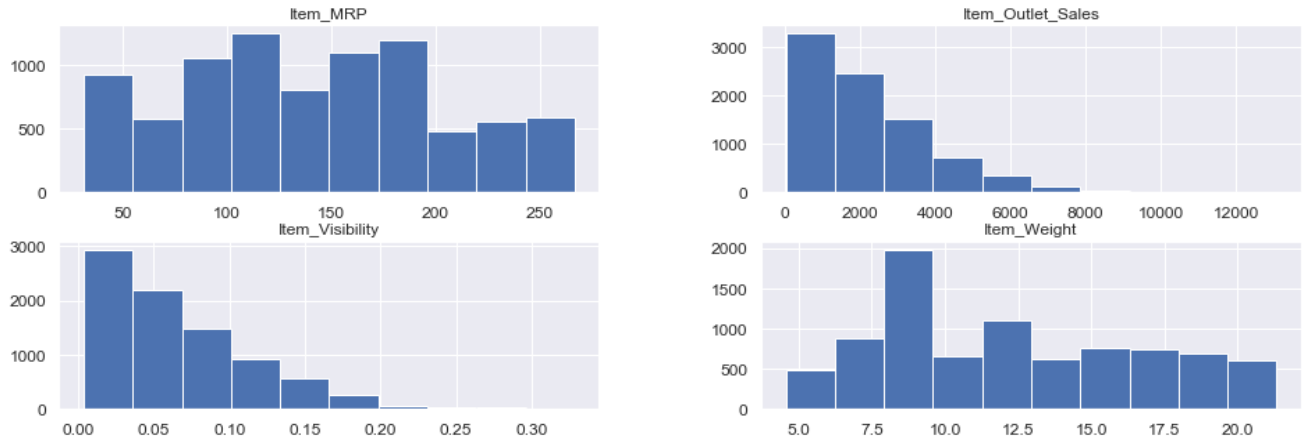

Figure 2. Histogram plot for Item MRP, Item Weight and Item Visibility

Figure 3 shows an example of univariate analysis using count plot. This graph shows the counts for different outlets. It includes categorical value counts.

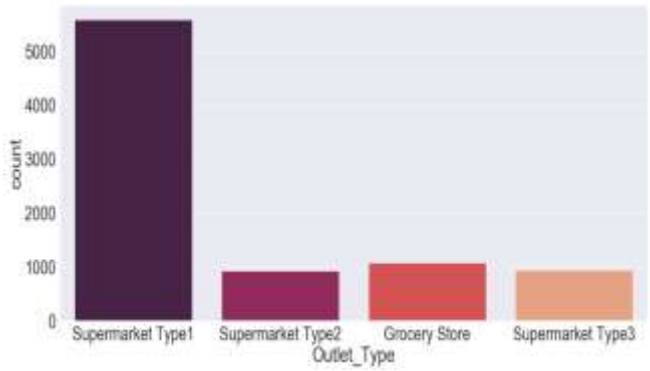

Figure 3. Count plot for types of outlet

Figure 4 shows the univariate analysis box plot distribution of the data. The box shows the quartiles of the data and the whiskers extend to show the rest of the distribution. It helps to detect outliers in the data set. There exists outliers for sales and visibility.

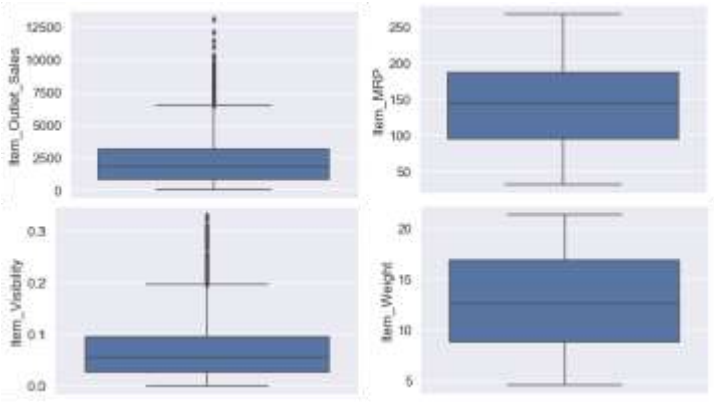

Figure 4. Box plot for Sales, Item MRP, Item Visibility and Item Weight

The other type of analysis Multivariate analysis. It refers to the statistical technique which is used to analyze data that consists of multiple variables. Univariate analysis looks at one variable at a time, multivariate analysis looks at two or more variables at a time to explore relationships. Multivariate EDA can be bivariate which consists of two variables, but often it will involves three or more variables. It is better to perform univariate EDA on each of the components of a multivariate EDA to study the trends in each variable and then perform the multivariate EDA. The main advantage of multivariate analysis is that you can study the relationship between variables. Multivariate non-graphical analysis shows the relationship between two or more variables in the form of either statistics or cross-tabulation. When there are multiple quantitative variables then pairwise correlation is assembled in a matrix. Figure 5 shows the correlation matrix for quantitative variables.

	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales
Item_Weight	1.000000	-0.038087	0.024206	-0.013181
Item_Visibility	-0.038087	1.000000	-0.001578	-0.129372
Item_MRP	0.024206	-0.001578	1.000000	0.567574
Outlet_Establishment_Year	0.188915	-0.075289	0.005020	-0.049135
Item_Outlet_Sales	-0.013181	-0.129372	0.567574	1.000000

Figure 5. Correlation between different variables

The correlation value ranges from +1 to -1. A negative value indicates a negative association and a positive value indicates a positive association for correlation. The closer the correlation value is to 1 the data points all fall in a straight line which gives a linear association. The closer the correlation value is to 0 gives a weaker linear association. Figure 6 shows the variable that has a positive correlation with the sales variable and the correlation value is greater than 0.5.

```
The variable having strong correlation with Item_Outlet_Sales:
Item_MRP            0.567574
Name: Item_Outlet_Sales, dtype: float64
```

Figure 6. Variable having correlation greater than 0.5 with sales

The above examples are non-graphical multivariate analysis. The following example includes a graphical multivariate analysis. Figure 7 is an example of bivariate analysis and it shows the relationship between two variables sales and outlet type. It is found that the grocery store has the least sales amongst all the other outlets, so some steps must be taken to help increase the sales at the grocery store. The products contributing to high sales include dairy products, soft drinks, and fruits and vegetables. After comparing the relationship between sales and outlet, we can discover trends or the relationship between sales and different products. Figure 8 is an example of a bivariate analysis using a scatter plot. From this graph, we discover that sales are high for fruits and vegetables and snack food. Box plot can also be used for bivariate analysis. Figure 9 shows the box plot for sales and products. This graph helps to understand outliers and what their values are. It shows the minimum, maximum, first and third quartile and the median for each product and its sale value. This will provide insights into whether the data set is symmetrical, how tightly they are grouped, and whether the data is skewed. Figure 5 was an example of non-graphical multivariate analysis this can be converted into graphical form using heatmaps. Heatmaps helps to study the correlation between different variables in a graphical format. Here the correlation values are represented in a matrix form using different colors. Figure 10 shows heatmap for multivariate analysis as it includes more than two variables.

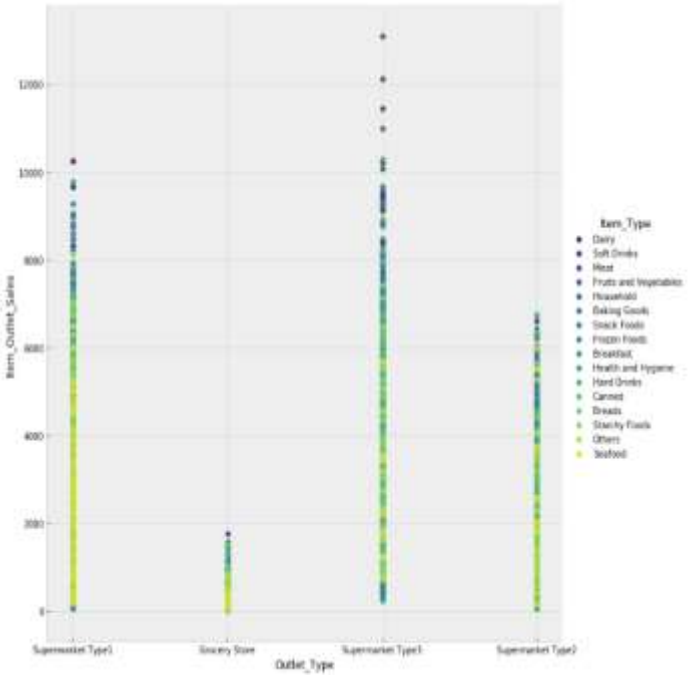

Figure 7. Bivariate analysis between Sales and Outlet Type

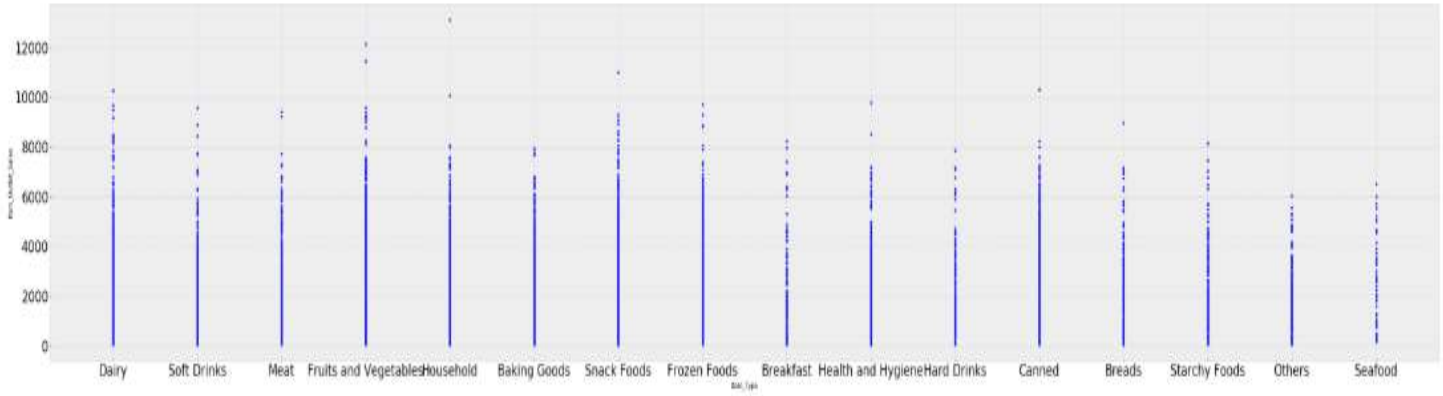

Figure 8. Scatter plot for Sales and Item Type

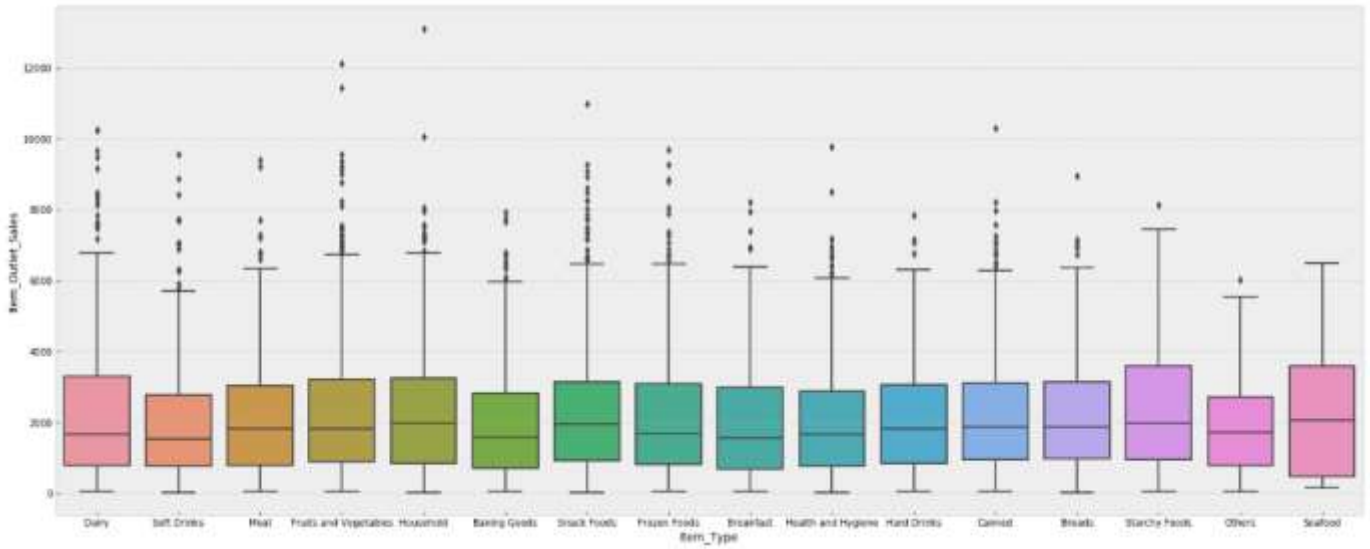

Figure 9. Box plot for Sales and Item Type

Figure 10. Heatmap for multivariate analysis

Positive values indicate a positive correlation between variables. The closer the value to 1 better the association. The only variable having a positive relationship with sales is the MRP of items. As there is a positive relationship between the two there exists a linear association between them. After knowing how these variables are correlated with each other, it is possible to see a visual graph for each pairs of variable using a pair plot. A pair plot is an effective tool for EDA. They allow us to see both distribution of a single variable and the relationship between two variables. They help to find trends, patterns, and relationships among variables. The diagonal axes include univariate distribution and are treated differently. Figure 11 shows the Pair plot for different variables. It also shows a regression line to show the trend between variables and whether they have a linear relationship. The trend line is linear for sales and MRP.

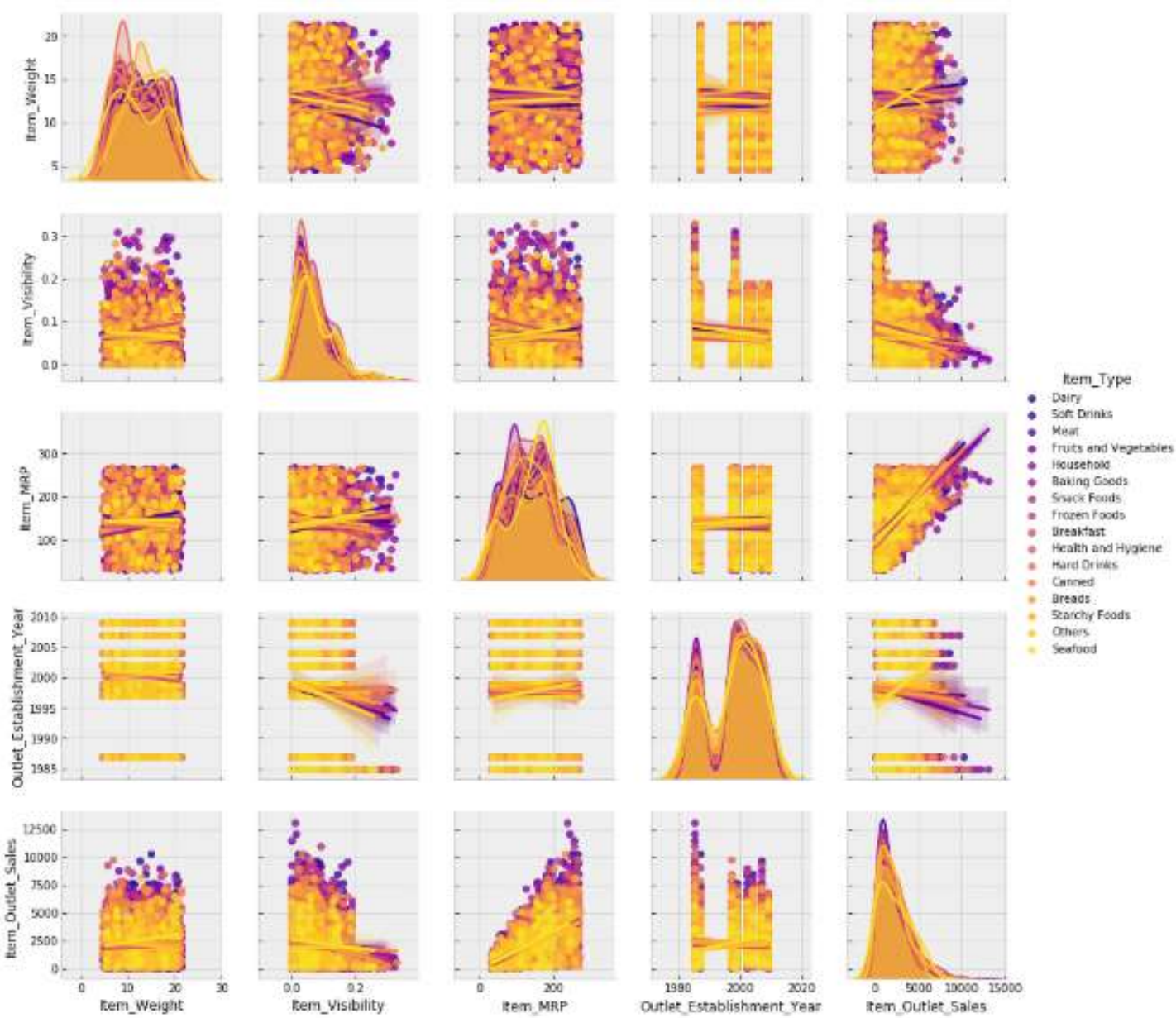

Figure 11. Pair plot for multivariate analysis

V. CONCLUSION

Using EDA within data governance has lots of advantages that help to bring more useful and actionable insights on data. It enhances the operational efficiency and ensures that the data is fit for purpose. It also helps to find the critical data elements within the data set. Using these insights it would be easy for data governance members to make an effective and informed decision that would help the business to achieve its objective. EDA helps the data governance team to save time and cost and gives more power to trust data governance decisions.

VI. REFERENCES

[1] A. Yamada and M. Peran, "Governance framework for enterprise analytics and data," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 3623-3631.

[2] T. He, S. Chen, L. Hao and J. Liu, "Quality Driven Judicial Data Governance," 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, 2019, pp. 66-70.

[3] D. Cheng, P. Schretlen, N. Kronenfeld, N. Bozowsky and W. Wright, "Tile based visual analytics for Twitter big data exploratory analysis," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, 2013, pp. 2-4.

[4] Cheong, Lai Kuan and Chang, Vanessa, "The Need for Data Governance: A Case Study" (2007). ACIS 2007 Proceedings. 100. http://aisel.aisnet.org/acis2007/100

[5] Abraham, Rene & Brocke, Jan vom & Schneider, Johannes. (2019). Data Governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management. 49. 10.1016/j.ijinfomgt.2019.07.008.

[6] Wikipedia contributors. (2020, April 18). Data governance. In Wikipedia, The Free Encyclopedia., from https://en.wikipedia.org/w/index.php?title=Data_governance&oldid=951669164

[7] Wikipedia contributors. (2020, April 16). Exploratory data analysis. In Wikipedia, The Free Encyclopedia., from https://en.wikipedia.org/w/index.php?title=Exploratory_data_analysis&oldid=951236179