19 November 2018
16:34

Bivariate Analysis

After looking at every feature individually, let's now do some bivariate analysis. Here we'll explore the independent variables with respect to the target variable. The objective is to discover hidden relationships between the independent variable and the target variable and use those findings in missing data imputation and feature engineering in the next module.

We will make use of **scatter plots** for the continuous or numeric variables and **violin plots** for the categorical variables.

```
train = combi[1:nrow(train)] # extracting train data from the combined data
```

Target Variable vs Independent Numerical Variables

Let's explore the numerical variables first.

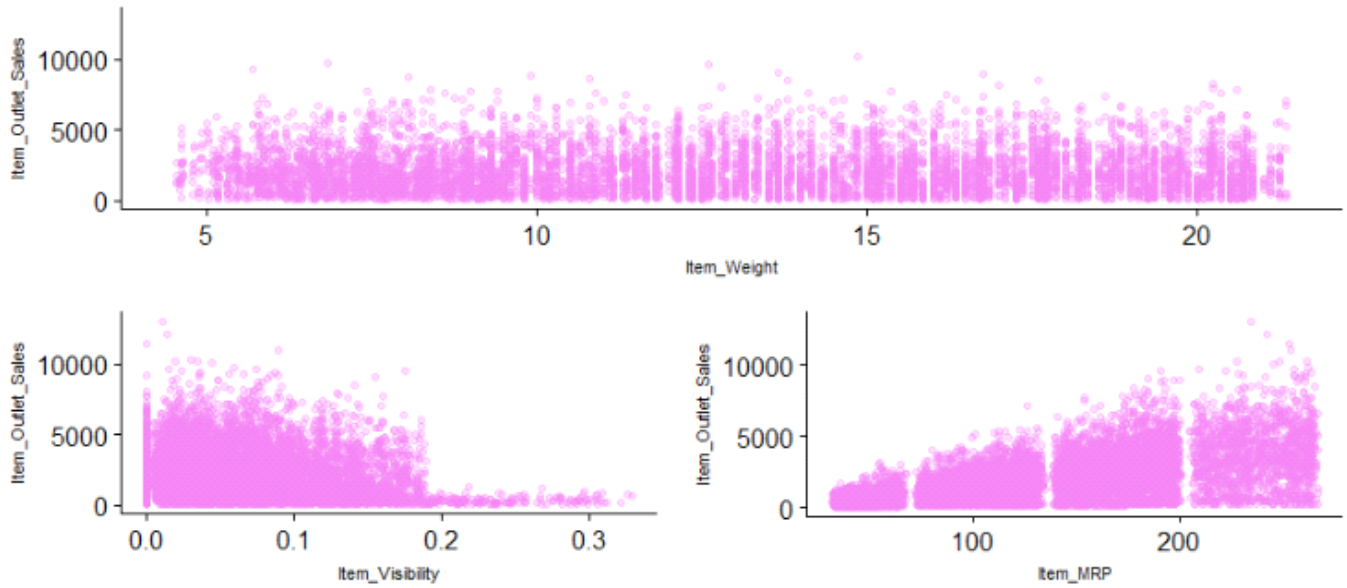
```
# Item_Weight vs Item_Outlet_Sales
p9 = ggplot(train) +
     geom_point(aes(Item_Weight, Item_Outlet_Sales), colour = "violet", alpha = 0.3) +
     theme(axis.title = element_text(size = 8.5))
```

```
# Item_Visibility vs Item_Outlet_Sales
p10 = ggplot(train) +
      geom_point(aes(Item_Visibility, Item_Outlet_Sales), colour = "violet", alpha = 0.3) +
      theme(axis.title = element_text(size = 8.5))
```

```
# Item_MRP vs Item_Outlet_Sales
p11 = ggplot(train) +
      geom_point(aes(Item_MRP, Item_Outlet_Sales), colour = "violet", alpha = 0.3) +
      theme(axis.title = element_text(size = 8.5))
```

```
second_row_2 = plot_grid(p10, p11, ncol = 2)
plot_grid(p9, second_row_2, nrow = 2)
```

```
Removed 1463 rows containing missing values (geom_point).
```



Observations

- Item_Outlet_Sales is spread well across the entire range of the Item_Weight without any obvious pattern.
- In Item_Visibility vs Item_Outlet_Sales, there is a string of points at Item_Visibility = 0.0 which seems strange as item visibility cannot be completely zero. We will take note of this issue and deal with it in the later stages.
- In the third plot of Item_MRP vs Item_Outlet_Sales, we can clearly see 4 segments of prices that can be used in feature engineering to create a new variable.

Target Variable vs Independent Categorical Variables

Now we'll visualise the categorical variables with respect to Item_Outlet_Sales. We will try to check the distribution of the target variable across all the categories of each of the categorical variable.

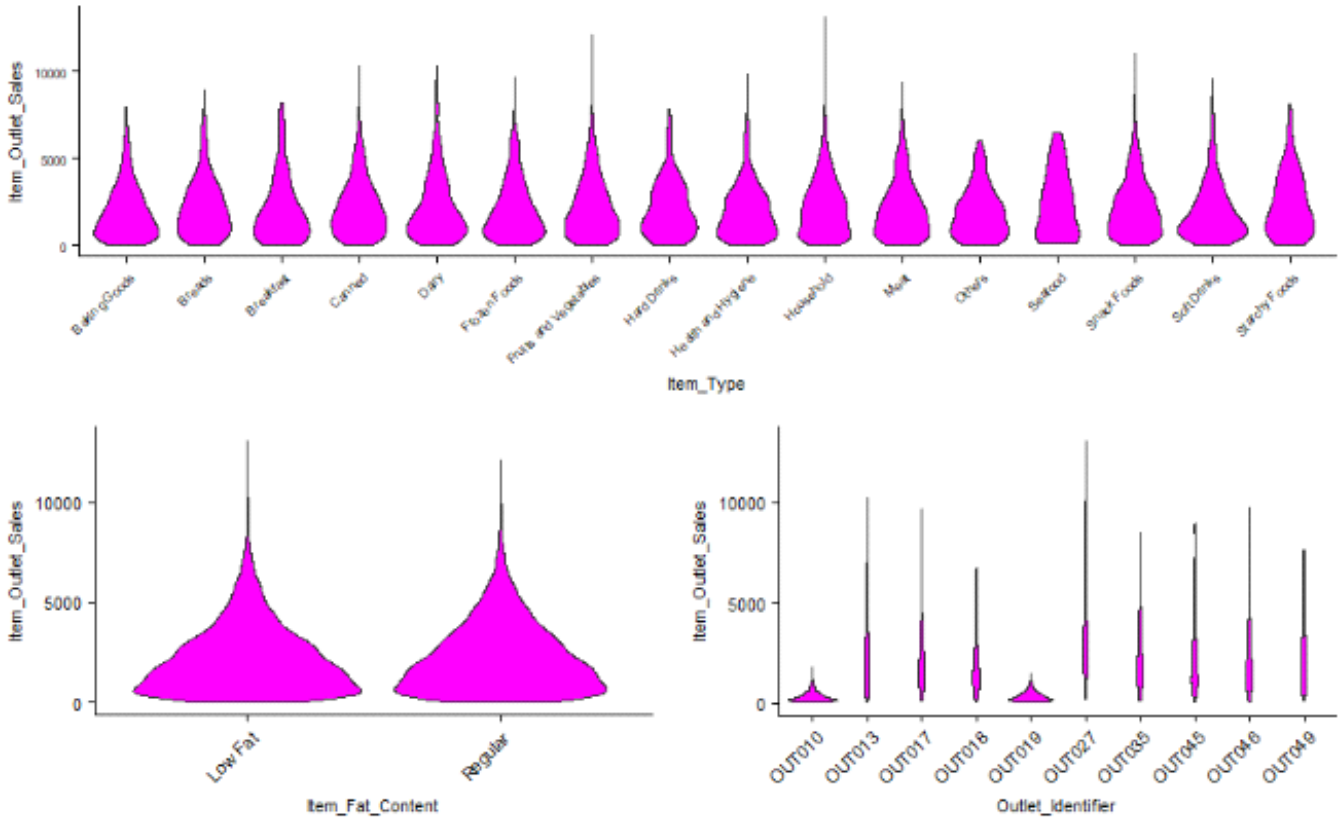
We could have used boxplots here, but instead we'll use the violin plots as they show the full distribution of the data. The width of a violin plot at a particular level indicates the concentration or density of data at that level. The height of a violin tells us about the range of the target variable values.

```
# Item_Type vs Item_Outlet_Sales
p12 = ggplot(train) +
      geom_violin(aes(Item_Type, Item_Outlet_Sales), fill = "magenta") +
      theme(axis.text.x = element_text(angle = 45, hjust = 1),
            axis.text = element_text(size = 6),
            axis.title = element_text(size = 8.5))
```

```
# Item_Fat_Content vs Item_Outlet_Sales
p13 = ggplot(train) +
      geom_violin(aes(Item_Fat_Content, Item_Outlet_Sales), fill = "magenta") +
      theme(axis.text.x = element_text(angle = 45, hjust = 1),
            axis.text = element_text(size = 8),
            axis.title = element_text(size = 8.5))
```

```
# Outlet_Identifier vs Item_Outlet_Sales
p14 = ggplot(train) +
      geom_violin(aes(Outlet_Identifier, Item_Outlet_Sales), fill = "magenta") +
      theme(axis.text.x = element_text(angle = 45, hjust = 1),
            axis.text = element_text(size = 8),
            axis.title = element_text(size = 8.5))
```

```
second_row_3 = plot_grid(p13, p14, ncol = 2)
plot_grid(p12, second_row_3, ncol = 1)
```

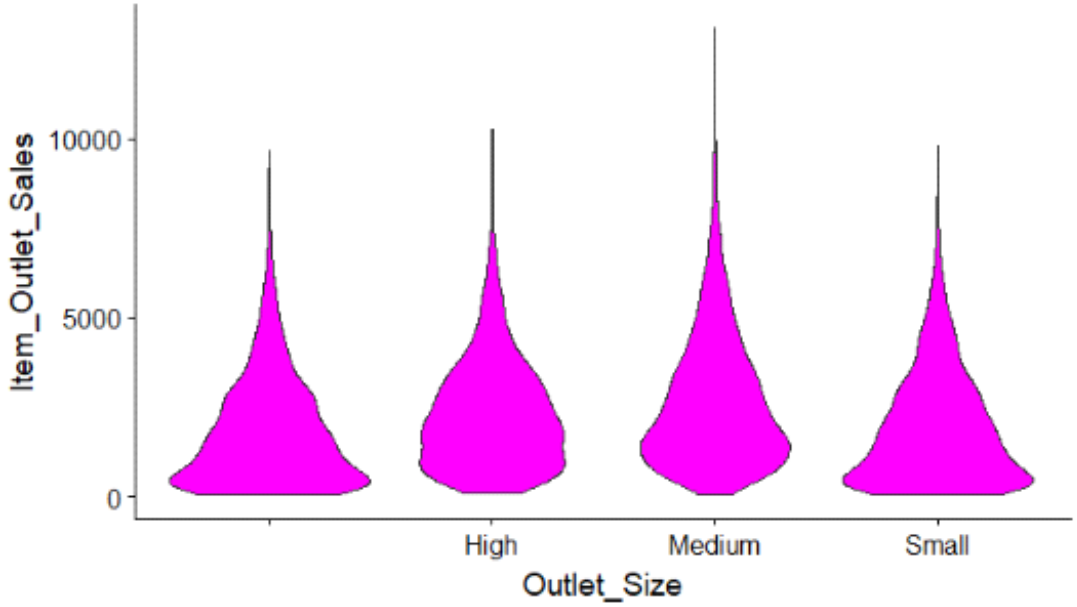


Observations

- Distribution of Item_Outlet_Sales across the categories of Item_Type is not very distinct and same is the case with Item_Fat_Content.
- The distribution for OUT010 and OUT019 categories of Outlet_Identifier are quite similar and very much different from the rest of the categories of Outlet_Identifier.

In the univariate analysis, we came to know about the empty values in Outlet_Size variable. Let's check the distribution of the target variable across Outlet_Size.

```
ggplot(train) + geom_violin(aes(Outlet_Size, Item_Outlet_Sales), fill = "magenta")
```

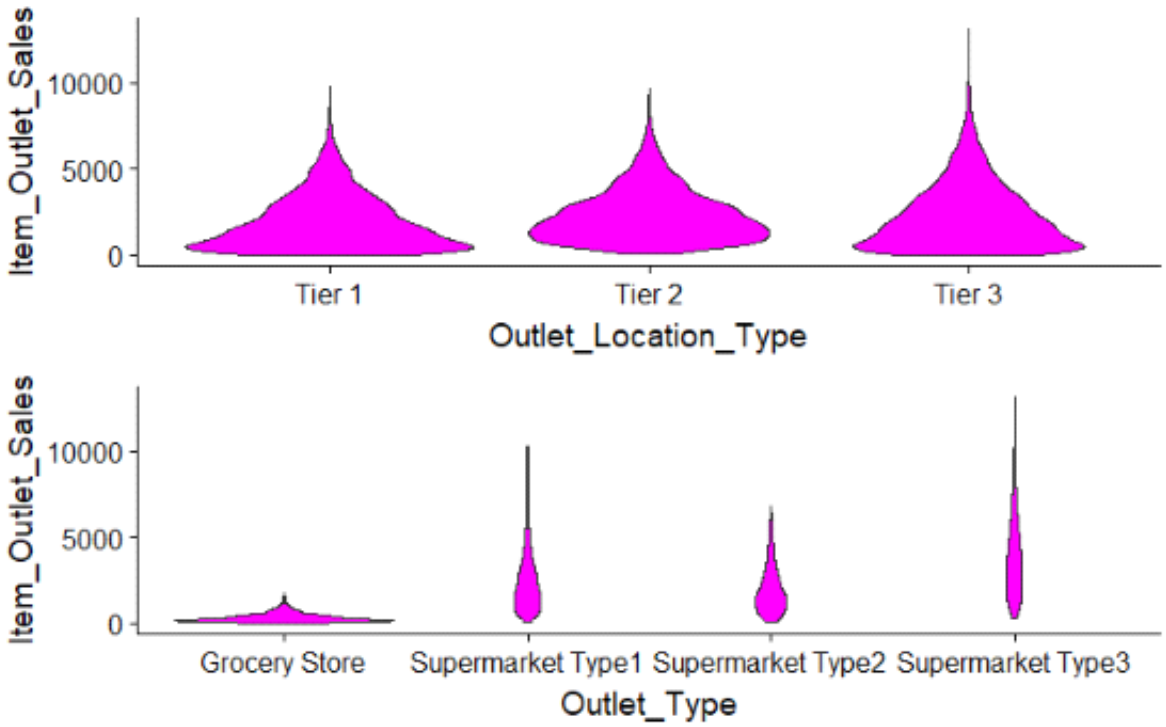


The distribution of 'Small' Outlet_Size is almost identical to the distribution of the blank category (first vioin) of Outlet_Size. So, we can substitute the blanks in Outlet_Size with 'Small'.

Please note that this is not the only way to impute missing values, but for the time being we will go ahead and impute the missing values with 'Small'.

Let's examine the remaining variables.

```
p15 = ggplot(train) + geom_violin(aes(Outlet_Location_Type, Item_Outlet_Sales), fill = "magenta")
p16 = ggplot(train) + geom_violin(aes(Outlet_Type, Item_Outlet_Sales), fill = "magenta")
plot_grid(p15, p16, ncol = 1)
```



Observations

- Tier 1 and Tier 3 locations of Outlet_Location_Type look similar.
- In the Outlet_Type plot, Grocery Store has most of its data points around the lower sales values as compared to the other categories.

These are the kind of insights that we can extract by visualizing our data. Hence, data visualization should be an important part of any kind data analysis.