Exploratory Data Analysis using R

Objective:

Dataset is of a fictional company found from the internet, and my aim is to try and see where the company is spending most of it's money on. I have used the R language for this, faceting over multiple variables to see what kind of relationships exist between the variables. I have also found correlations and then done multiple-regression to see how powerful the model is.

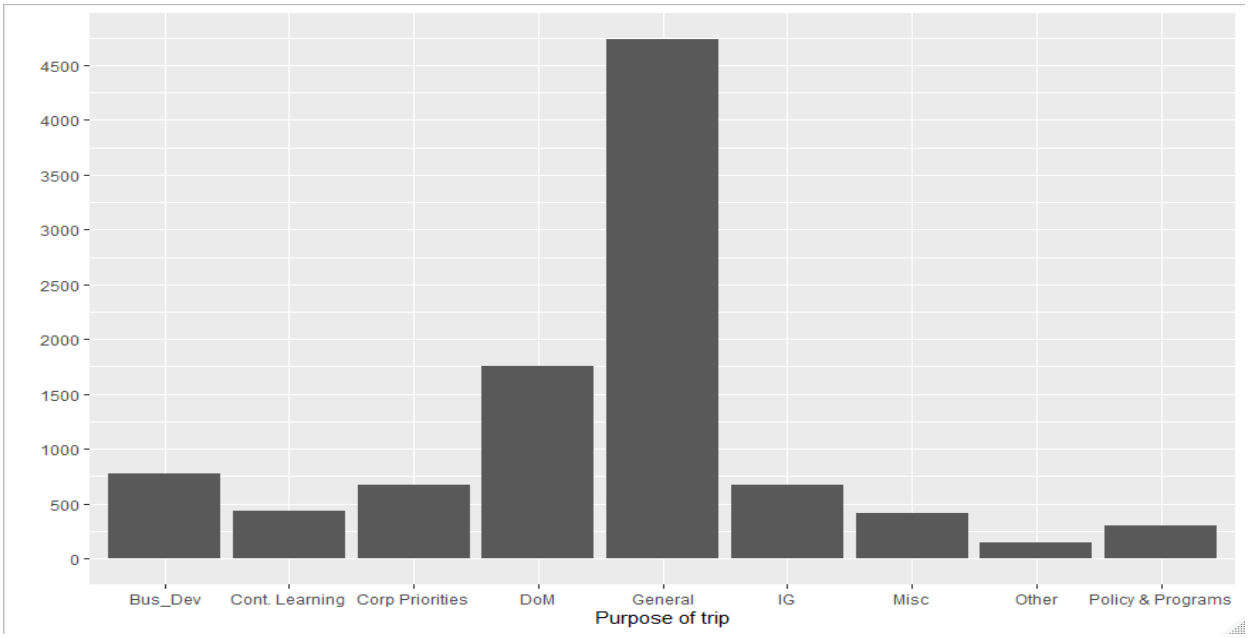
The dataset is not very good, and hence, the objective is not so much on the analysis. It is more for my GitHub profile and to show that I am comfortable in writing R code. Check R_EDA_code.R

Executive Summary:

Employees made most of their trips to Ottawa, Ontario and Quebec and most for 'General' and Delivery of Mandate Purposes (1). I am working with 4 variables to explain the rise in 'Total' costs: 'Purpose of trip', 'Month' 'Region', 'Nights'. After running the regression, even though Region and Month are statistically significant, the model has an R^2 of only 2% and hence these variables don't explain the model well, showing the lack of credibility of the dataset (2). We see that 'General' is dense with costs, especially under the $500 range. Same goes for Delivery of Mandate and Misc (3). We do see that in NAT (Ottawa), 'General' makes up a large proportion of expenses. Same goes for Misc expenses in NAT (Ottawa) (4). In each of the provinces, DoM and General expenses make up a large proportion of the expenses. The graphs become more dense in seasons 2 and 4, signalling more of these expenses being incurred in NAT and ONT during summer and fall months (5). An example is that a lot of General expenses such as training costs, hosting events and outdoor events take place in NAT (Ottawa) and the rest of ONT during the summer months.

Details:

1- Working with one variable: I have started with some data cleaning, focusing on the 'Purpose of trip' variable. Most trips are 'General' which consists of training, housing projects and so on. Finding total costs for each of these "Purpose of trip", I see that they are positively skewed since the Mean is greater than the Median for each.

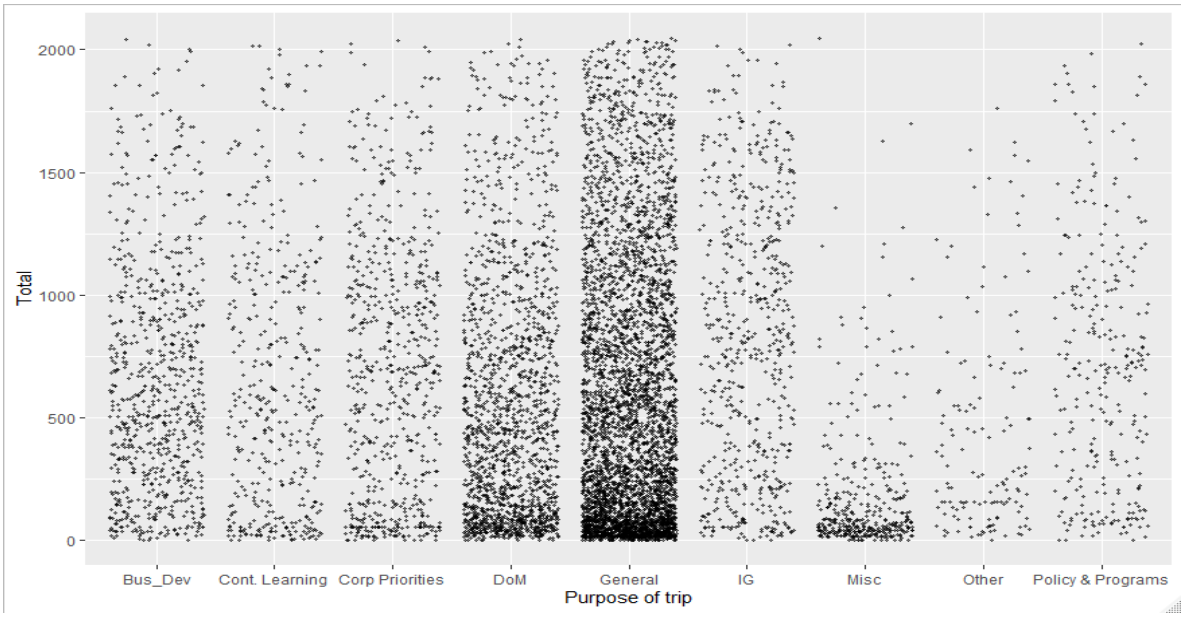


2- Correlation and regression:
I find the correlation matrix for 4 variables: 'Purpose of trip', 'Total', 'Month' 'Region', 'Nights' and 'Total'. Even though Total (response variable) is not correlated with the other 4 explanatory variables, I have still conducted a regression to find if they are associated with 'Total' in any way.

Based on a null-hypothesis that there is no relationship between the X and Y variables and looking at a 10% significance level, 'Region' and 'Month' have p-values < 10% and hence we can reject the null hypothesis. We can conclude that Region and Month are associated with an increase in Total costs. The 'Purpose of trip' has a p-value > 10% so we cant say the same for this.

However, the regression model has an R^2 value of 2% which is very poor and hence our explanatory variables can not be used to explain the model. However, based on intuition, none of the other variables from the dataset can be really used to predict costs, and that is why it is not a very good dataset.

3- Working with 2 variables: ‘Purpose of trip’ and ‘Total’: The graph had a lot of over-plotting so I added some jitter to solve the issue. We see that ‘General’ is dense with costs, especially under the $500 range. Same goes for DoM and Misc.



4- Working with 3 variables: ‘Purpose of trip’ and ‘Total’ and facet over ‘Region’. After adjusting for over-plotting, even though its not clear to see patterns, we do see that in NAT (Ottawa), ‘General’ makes up a large proportion of expenses. Same goes for Misc expenses in NAT (Ottawa).

5- Working with 4 variables: 'Purpose of trip' and 'Total' and facet over 'Region' and the 4 seasons in a year. I made a new variable for the 4 seasons based on the 'Month' variable. For instance, Months 1,2 and 3 would be in season 1. Adjusting for over-plotting, we see that in each of the provinces, DoM and General expenses make up a large proportion of the expenses. We see that the graphs become more dense in seasons 2 and 4, signalling more travelling and more hospitality expenditures during the summer and fall months.