**MIS 584 Business Intelligence**

Class 3: In-class data analysis exercise

Worcester Polytechnic Institute, 100 Institute Road.

*Contact: dasanghavi@wpi.edu,*[*kmpinto@wpi.edu*](mailto:kmpinto@wpi.edu)

1. **How many customers do we have?**

By taking customer Id as a feature to identify unique customers, we have 1180 customers.

Also, we analyzed the count of number of customer on the basis of State as that can be useful for the manager to make decisions.

State wise customer analysis

1. **Who are our customers? Specifically, characterize them in terms of their:**
   * **Gender**

There are 4 gender of customers. Also, there are 504 missing values for gender in customer database. Hence, after ignoring those rows the count can be determined from below graph:

* + **Age**

There are 244 rows with blank and noise data. After filtering those values, we calculated age using birth date and have discretized it in the bin of 20 to have better insights. We calculated the age by the following formula –

**Age = Num((today()-birthday)/365)**

Below is the distribution:

* + **Income**

We have discretized the income attribute into bins of 25000. There are unknown or missing values as well which are tagged as unknown.

* + **Geographic location**

As done above:

* **Any relationships among these**

After filtering Unknown rows or missing values, we can have an analogy between gender and age:

We can see that, Females don’t have customers of age 0-20. Also, X type don’t have enough customers and distribution of 0-20 age group is really less compare to other age groups.

Also, as we know, FL and NY has large amount of customers, we can do further analysis on them.

We can see that only with certain age group and income group are eligible to be our customers. So, manager can go in that direction for further profit.

*Overall, who are our typical customers?*  
People with age range of 21-120, gender types if F or M, Income range of 25000-150000 and from state like FL and NY are usually our customers.

1. **What are the quality problems with our customer data?**

* There are lot of missing value in most of the columns like Gender, Age, Income, Birthdate
* Also, there is noise data such as in date column, we have ‘#######’ as entry
* We have column such as country, middle name, Region, salutation as entirely missing