1. Determine the statistical variable types for each of the 37 columns in your DataMart (MYSTORE table) and put these into your summary document.

|  |  |
| --- | --- |
| **Column Name** | **Variable Data Type** |
| Visit\_Nbr | Continuous |
| Store\_Nbr | Continuous |
| Register\_Nbr | Continuous |
| Card\_Holder\_Nbr | Continuous |
| Membership\_Nbr | Continuous |
| Member\_Code | Categorical |
| Tender\_Type | Continuous |
| Tender\_Amt | Continuous |
| Sales\_Tax\_Amt | Numeric |
| Total\_Visit\_Amt | Continuous |
| Transaction\_Date | Categorical |
| Transaction\_Time | Continuous |
| Refund\_Code | Categorical |
| Tot\_Unit\_Cost | Continuous |
| Tot\_Unique\_Itm\_Cnt | Continuous |
| Tot\_Scan\_Cnt | Continuous |
| MEM\_ZIP | Categorical |
| MEMBER\_TYPE | Categorical |
| MEMBER\_STATUS\_CD | Categorical |
| ISSUING\_CLUB\_NBR | Continuous |
| RENEWAL\_DATE | Categorical |
| JOIN\_DATE | Categorical |
| Store\_Name | Categorical |
| STO\_ZIP | Continuous |
| Item\_Quantity | Continuous |
| Total\_Scan\_Amount | Continuous |
| Unit\_Cost\_Amount | Continuous |
| Unit\_Retail\_Amount | Continuous |
| Tax\_Collect\_Code | Categorical |
| Item\_Nbr | Continuous |
| Category\_Nbr | Continuous |
| Sub\_Category\_Nbr | Continuous |
| Primary\_Desc | Categorical |
| Secondary\_Desc | Categorical |
| Create\_Date | Categorical |
| Effective\_Date | Categorical |
| Sub\_category\_desc | Continuous |

1. n/a all results will be in the TWM Project in an analysis called “Step 2 Histogram ColumnName”
2. n/a all results will be in the TWM Project in an analysis called “Step 3 Frequency”

Top tender types (type: count): 1(Check): 130953, 0(Cash): 57024, 3(Discover): 17045

1. Which day of the week has the most items scanned? What is that total for that day for the month (you may have to break out the calculator on this one)?

Saturdays have the highest amount of total sales: 56,283

The day (also a Saturday) with the most sales is 1/15/2000 with a total of 15,636.

1. Which day of the week has the fewest items scanned? What is that total for that day for the month (you may have to break out the calculator on this one)? Include this answer in your summary doc.

Wednesdays have the least amount of total sales: 20,033

The day (also a Wednesday) with the least sales is also 1/12/2000 with total 3,216.

Note for 4 and 5: just because the total amount of sales on a day of the week for a month is higher than other days, it does not mean that day will have the most sales in one given week of observed month.

1. n/a all results will be in the TWM Project in an analysis called "Step 6 Frequency Issuing Club Nbr"
2. n/a all results will be in the TWM Project in an analysis called "Step 7 Frequency MemZip"

Where are most club members located?

Top 3 zip codes are(zip code(city, state): count):

54956(Neenah, Wisconsin): 21380 this is where most of the members are located.

55692(n/a): 16,179

54914(Appleton, Wisconsin): 16,061

Note: that the zip codes in the provided USPS link did not return a city, state.

1. Create Data Explorer and Frequency analysis blocks called "Step 8 Data Explorer" and "Step 8 Statistics" in order to determine the mean, median and standard deviations for the following variables and in your summary doc, provide a brief description of what this data tells you. If you find something (trust me, there is something odd in one of these columns) that is throwing off your statistics, what is it and how is it affecting your stats?
   1. Total\_Unique\_Item\_Count
   2. Total\_Visit\_Amt

225,305 total rows being analyzed. Median lies between row 112,652 and 112,653.

**Total\_Visit\_Amt:**

“Strange”: The MINIMUM amount a customer **card holder** visited the store was -1584.98.

Median: 115.96

Mean: 155.52

SD: 1419.27

The max amount of visits was 300,000.

Total\_Visit\_Amt represents the total amount of times a Membership\_Nbr visited MYSTORE.

The customer must have been sharing their card with others so this throws the data analytics off. Each card can represent more than one person and should never be less than one.

**Unique\_Itm\_Cnt:**

Median: 12.00

Mean: 14.27

SD: 11.55

Max: 141

**Explanation of Data**

Drilling the data I found that the Tot\_Unique\_Itm\_Cnt is always greater than or equal to the Tot\_Scan\_Cnt. Joined with the Transaction\_time attribute, I finally noticed that the total quantity and total scan count are the result of a transaction (implying each transaction is also the same card holder, Visit\_nbr, etc). Thus it is the total amount of unique items the member bought at that transaction time (there can be multiple transactions in a day).

More strange things I noticed were negative Item\_Quantity and Total\_Visit\_Amt. A negative total visit amount can result in a negative skew thus pulling the mean down. Total Unique amount of items is not affected by a negative amount of Item\_Quantity. I saw nothing that would hinder the median, mean, SD results of Tot\_Unique\_Itm\_Cnt.