Data preprocess report – Group 3

Introduction:

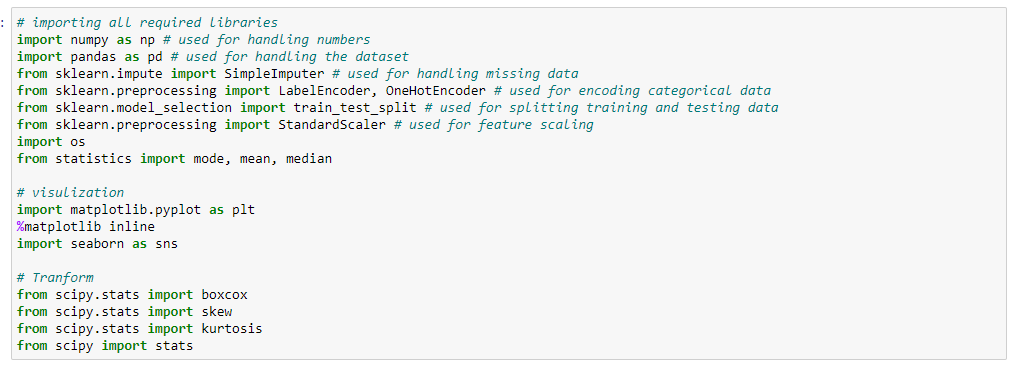
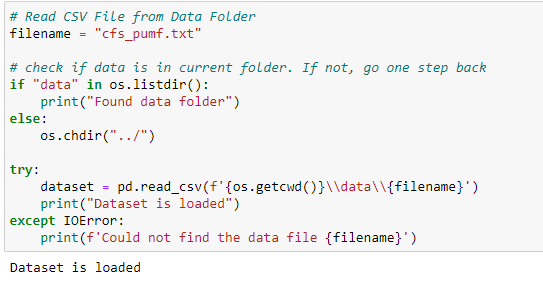
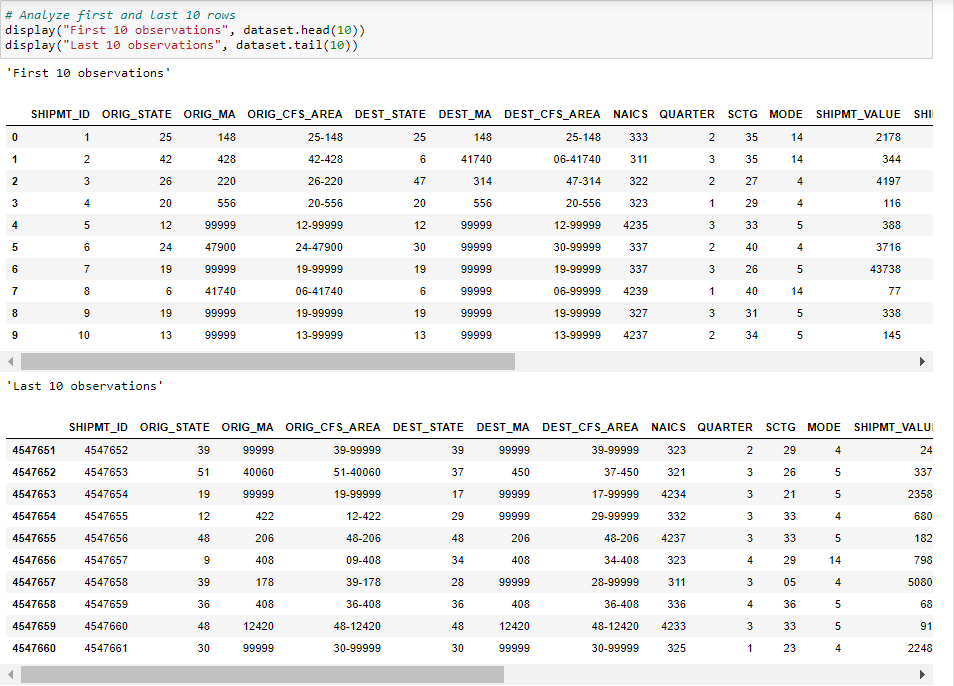
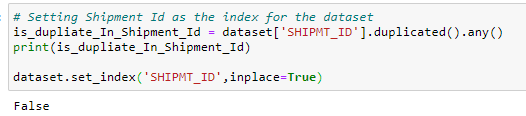
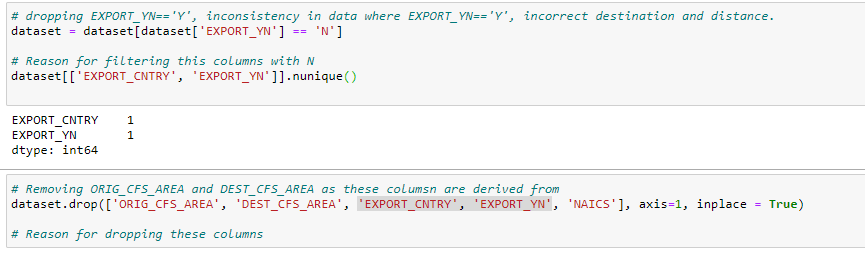
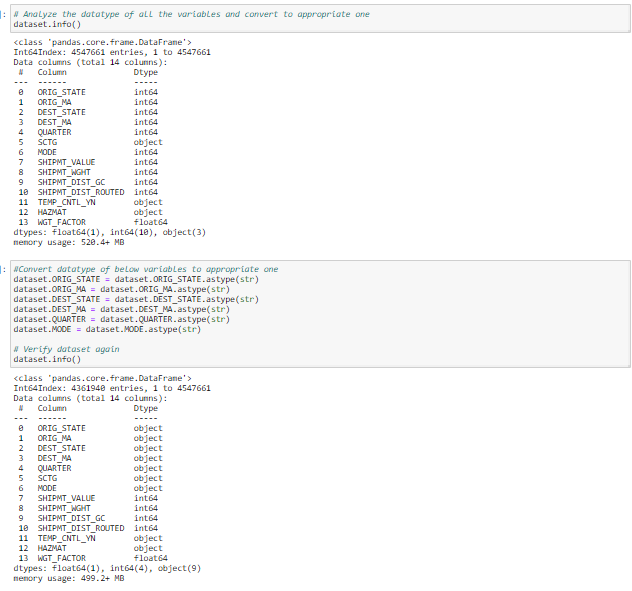
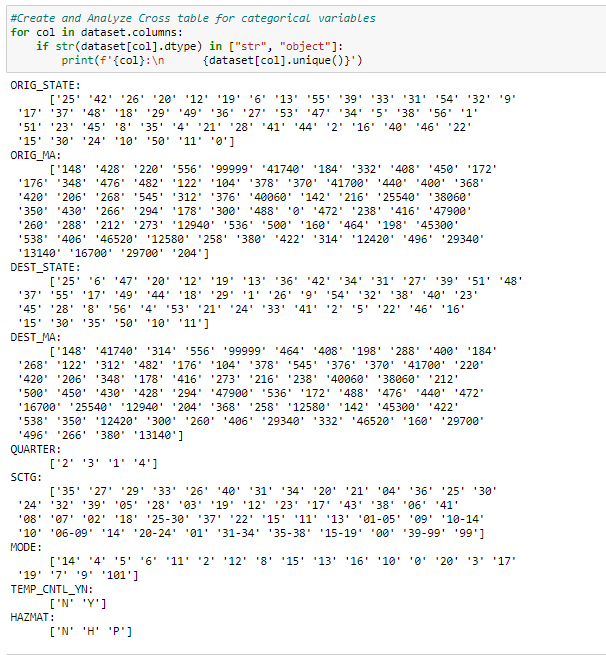
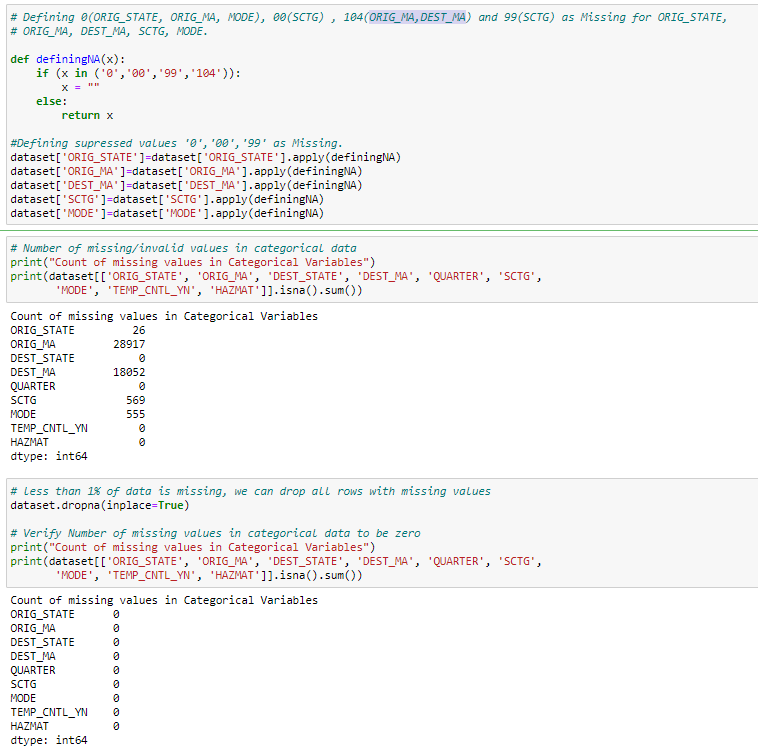
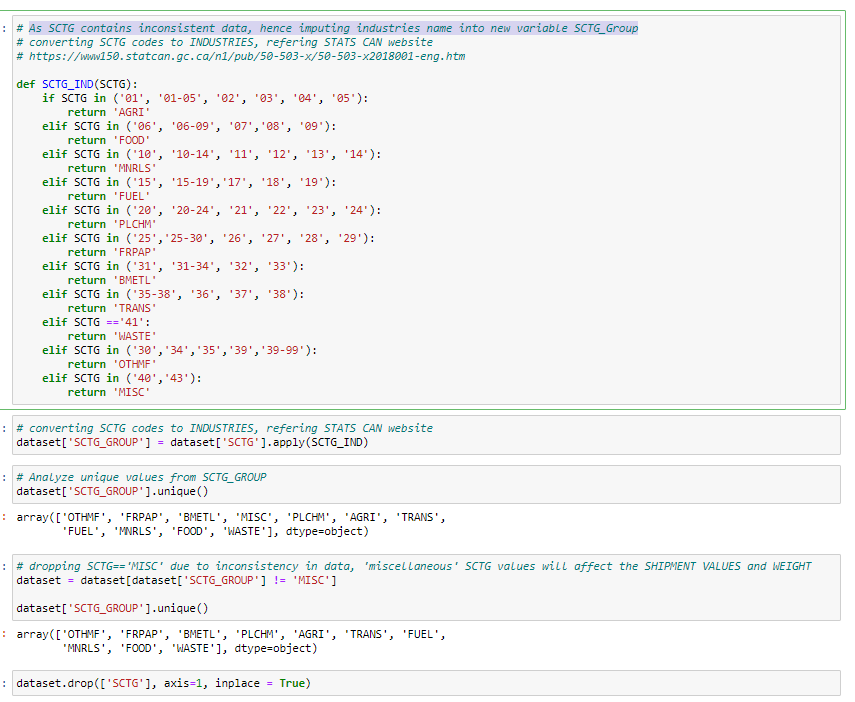
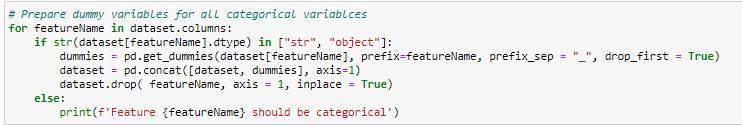
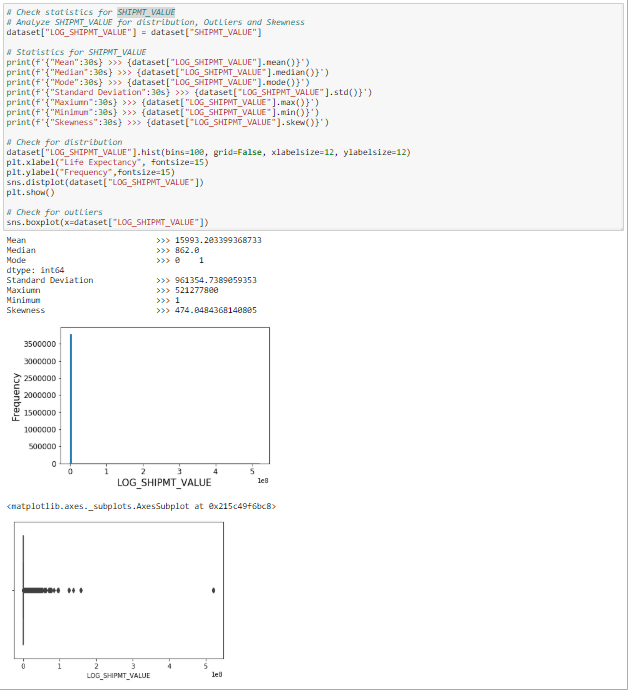
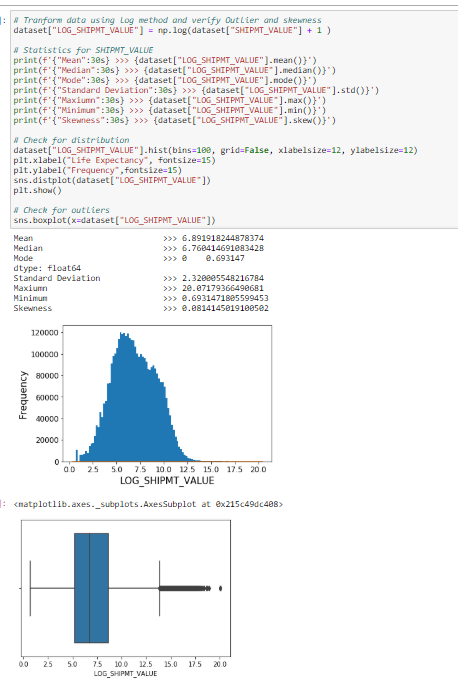
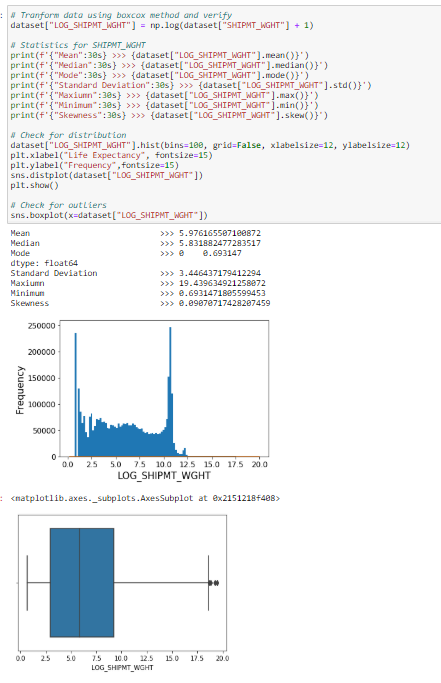
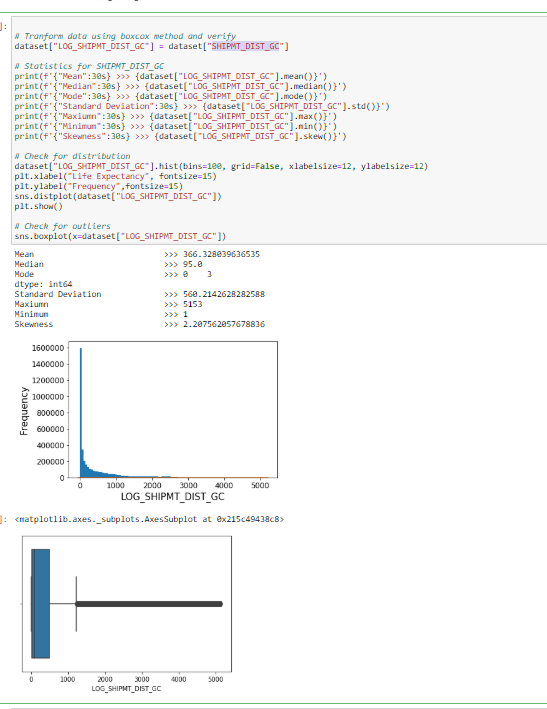
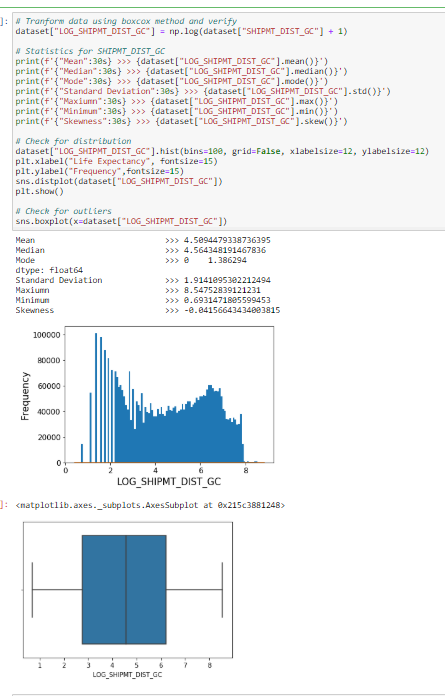
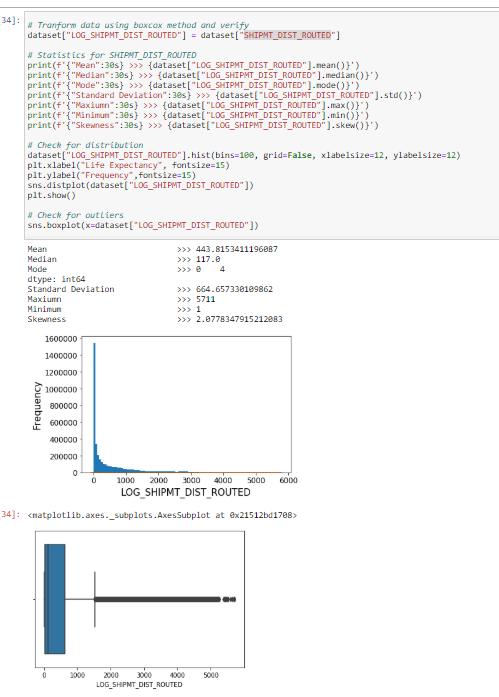
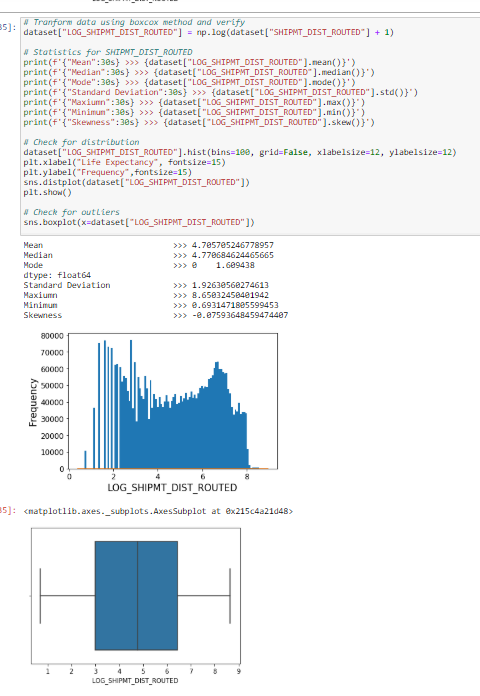
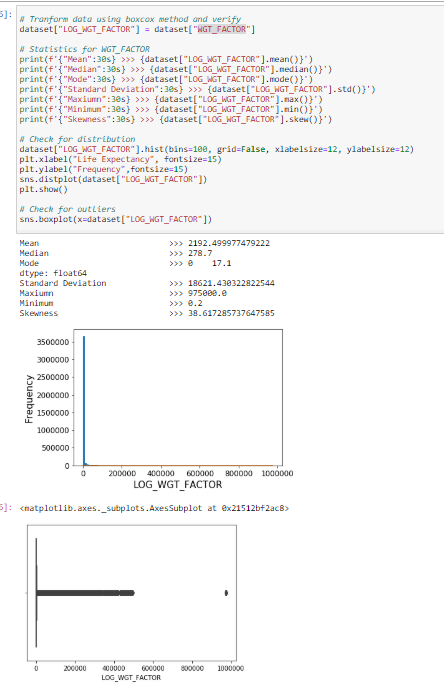
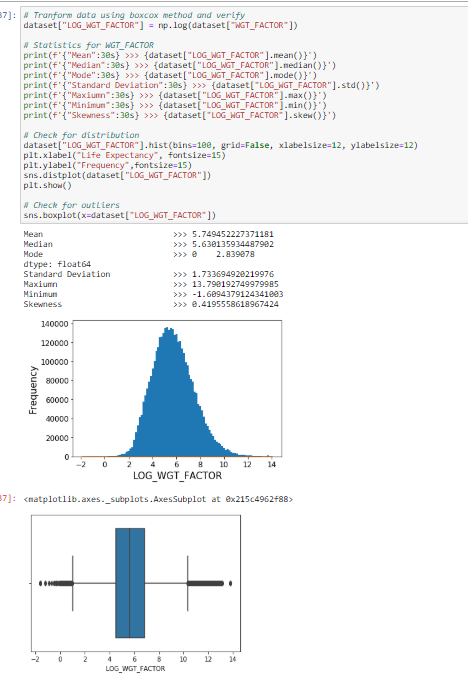
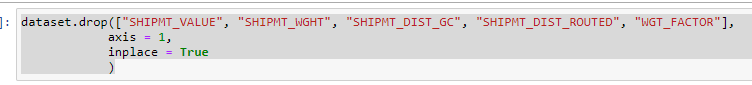
Before Processing Data:

* Variables:
  + Categorical: 14
  + Quantitative: 6
* Observations: 4547661
* Data Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| **CFS PUM File Data Dictionary** |  |  |  |
| **Field** | **Description** | **Type** | **Length** |
| **SHIPMT\_ID** | Shipment identifier | NUM | 7 |
| **ORIG\_STATE** | FIPS state code of shipment origin | CHAR | 2 |
| **ORIG\_MA** | Metro area of shipment origin | CHAR | 5 |
| **ORIG\_CFS\_AREA** | CFS Area of shipment origin | CHAR | 8 |
| **DEST\_STATE** | FIPS state code of shipment destination | CHAR | 2 |
| **DEST\_MA** | Metro area of shipment destination | CHAR | 5 |
| **DEST\_CFS\_AREA** | CFS Area of shipment destination | CHAR | 8 |
| **NAICS** | Industry classification of shipper | CHAR | 6 |
| **QUARTER** | Quarter in which the shipment occurred | CHAR | 1 |
| **SCTG** | 2-digit SCTG Commodity Code of the shipment | CHAR | 5 |
| **MODE** | Mode of transportation of the shipment | CHAR | 2 |
| **SHIPMT\_VALUE** | Value of the shipment in dollars | NUM | 8 |
| **SHIPMT\_WGHT** | Weight of the shipment in pounds | NUM | 8 |
| **SHIPMT\_DIST\_GC** | Great circle distance between ship-ment origin and destination (in miles) | NUM | 8 |
| **SHIPMT\_DIST\_ROUTED** | Routed distance between shipment origin and destination (in miles) | NUM | 8 |
| **TEMP\_CNTL\_YN** | Temperature controlled shipment - Yes or No | CHAR | 1 |
| **EXPORT\_YN** | Export shipment - Yes or No | CHAR | 1 |
| **EXPORT\_CNTRY** | Export final destination | CHAR | 1 |
|  |  |  |  |
|  |  |  |  |
| **HAZMAT** | Hazardous material (HAZMAT) code | CHAR | 1 |
|  |  |  |  |
|  |  |  |  |
| **WGT\_FACTOR** | Shipment tabulation weighting factor | NUM | 8 |

After Processing Data:

* Variables:
  + Index: 1
  + Numerical: 276
    - Continuous: 5
    - Dummy: 271
  + Target Variable: SHIPMT\_VALUE
  + Dropped: 5
* # Observations: 3,792,057

1. Libraries imported for data pre-processing:  
   
2. Importing the dataset:  
   
3. Snapshot for first and last 10 records of the dataset  
   
4. We found that SHIPMT\_ID is an index and does not contains any missing or duplicate values, so we marked this column as index.  
   
5. From the EXPORT\_YN, we removed the observations with Export as Y because of incorrect destination and distance.
6. Team decided to remove **ORIG\_CFS\_AREA and DEST\_CFS\_AREA** becausethese columns are derived from **ORIG\_STATE, ORIG\_MA and DEST\_STATE, DEST\_MA** by concatenation
7. As we already filtered the records from **EXPORT\_YN** so **EXPORT\_YN** and **EXPORT\_CNTRY** are of no use because there is only one single value. i.e. “N”  
   
8. For categorical features, datatype was imputed as “**int”** whereas it should be either “**str”** or “**object**”. So, we manually updated the datatype.  
   
9. Preprocessing for Categorical Variables
   1. Analyzing unique values:  
      
   2. We found few missing/invalid values such as (0, 00, 99, 104) in ORIG\_STATE, ORIG\_MA, MODE, SCTG, DEST\_MA as per data dictionary. So, we removed those observations which stands for less than 1% of overall dataset.  
      
   3. As SCTG contains inconsistent data such as few observations contain group number such as 01-05, 06-09 whereas other contains individual industry codes. Hence, we decided to gather industries name from “**STATS Canada”** andderived a new variable named as SCTG\_Group. Also, there were few codes such as 40 and 43 where industry name was not mentioned. So, we grouped them into one and removed them.  
      
   4. For all the categorical variables, we created dummy variables as we are planning to perform regression analysis.  
      
10. **Preprocessing for Quantitative Variables**
    1. Analyzed “**SHIPMT\_VALUE**”, “**SHIPMT\_WGHT** “, “**SHIPMT\_DIST\_GC**”, “**SHIPMT\_DIST\_** **ROUTED**” and “**WGT\_FACTOR**” features and found that there is too much noise all of these. We also found that data is highly skewed and there are many outliers. We also observe that the variance if quite high. So, we decided to normalize the data and transform it. We tried various techniques such as **boxcox, cube, exponential and log**. After analyzing all, we found that **log(x + 1)** is the best possible formula to transform the data. Though it removed the skewness, yet there are many outliers which cannot be removed as they are the important datapoints.  
       Before Transformation  
         
       After Transformation:  
         
       Shipment Weight  
         
         
       SHIPMT\_DIST\_GC  
         
         
       SHIPMT\_DIST\_ROUTED  
         
         
       WGT\_FACTOR  
         
       
    2. After the log transformation, we removed the columns with raw data and kept the transformed features.  
       
    3. We did not find any missing/incorrect/inconsistent data in the data. For example, we check the value and weight for value which either 0 or less than 0 as these could be invalid.
11. Snapshot of data after pre-processing:  
    