Additional Analysis

Task 1: You need to perform additional analysis on various variables and make a report.

* You might want to study which variables are highly correlated. If you find such variables you can suggest dimension reduction by dropping one of the variables.

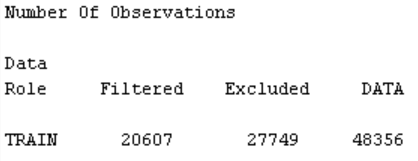
The following variables are highly correlated based a correlation of greater than 0.8. (<http://www.sfu.ca/~dsignori/buec333/lecture%2016.pdf>)

|  |  |  |
| --- | --- | --- |
| Variable 1 | Variable 2 | Correlation |
| days since last | months since last | 0.999975251 |
| months since last | days since last | 0.999975251 |
| avg $ net demand | total $ demand | 0.993953254 |
| total $ demand | avg $ net demand | 0.993953254 |
| avg $ demand | tot $ net demand | 0.953178097 |
| tot $ net demand | avg $ demand | 0.953178097 |
| tot units demand | total $ demand | 0.881179133 |
| total $ demand | tot units demand | 0.881179133 |
| avg $ net demand | tot units demand | 0.877361971 |
| tot units demand | avg $ net demand | 0.877361971 |
| lifetime orders | total $ demand | 0.815401727 |
| total $ demand | lifetime orders | 0.815401727 |
| avg $ net demand | lifetime orders | 0.812388543 |
| lifetime orders | avg $ net demand | 0.812388543 |
| lifetime orders | tot units demand | 0.804471837 |
| tot units demand | lifetime orders | 0.804471837 |

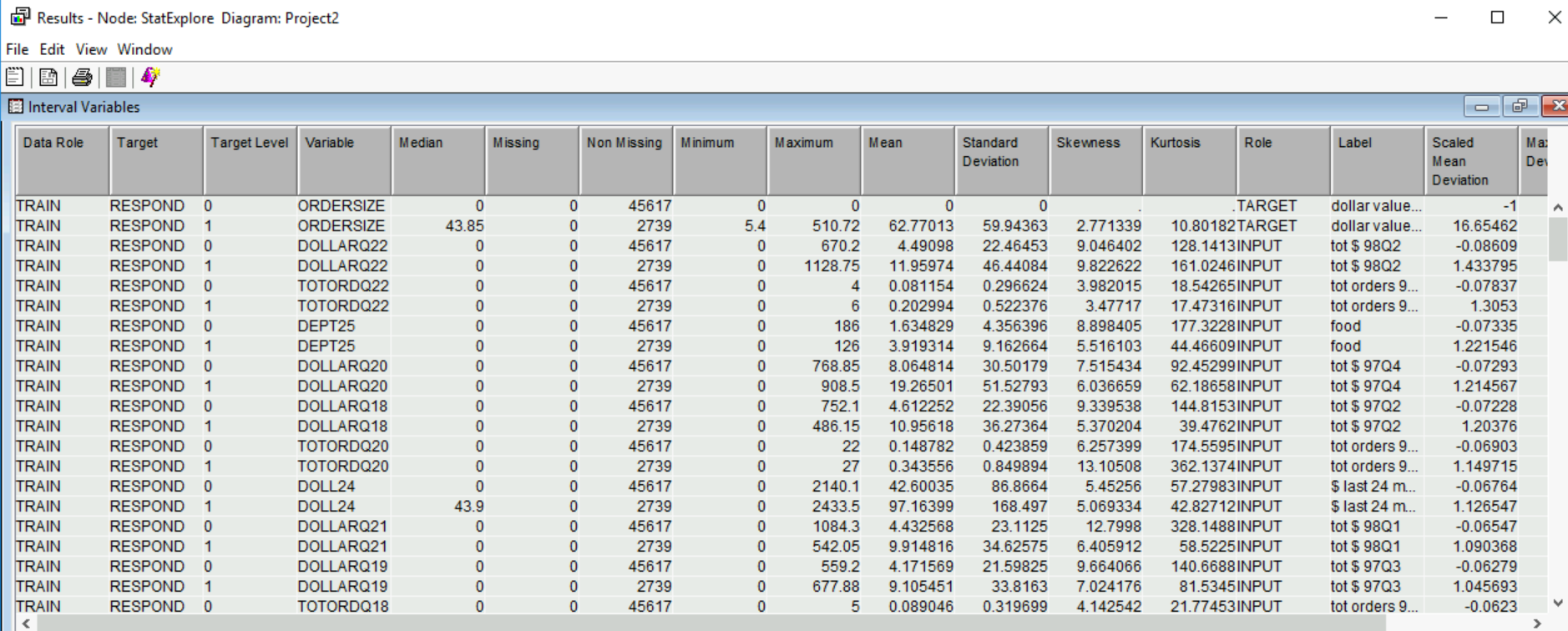
Highly correlated variables may overly skew data analysis in specific directions. One remedy is to drop one of each pair of highly correlated variables. In the CATALOG2010 data, we would drop “months since last” and keep “days since last” because days is a more precise measure. Between variables “avg $ net demand” and “total $ demand”, we would drop “avg $ net demand” because it is a comparison value and keep “total $ demand” because it is a final demand value rather than a value for comparison. For the same reason, we would keep “avg $ demand” instead of the comparison value of “tot $ net demand”. When reviewing “tot unit demand” and “total $ demand”, we have already indicated we will keep “total $ demand” so “tot unit demand” can be dropped. The next pair is “avg $ net demand” and “tot units demand”. We have already indicated that we will drop both. If we must retain one, it would be “tot units demand” because it represents a final concrete value. When comparing variables “lifetime orders” and “total $ demand”, we have already indicated that we will keep “total $ demand”. Between “avg $net demand” and “lifetime orders”, we can indicated that we will drop both of these. However, if we must retain one it will be “lifetime orders”. In comparing “lifetime orders” and “tot units demand”, we will drop “tot units demand” because it is more highly correlated with another variable – “total $ demand”.

* You can study in if there are outliers in your variables

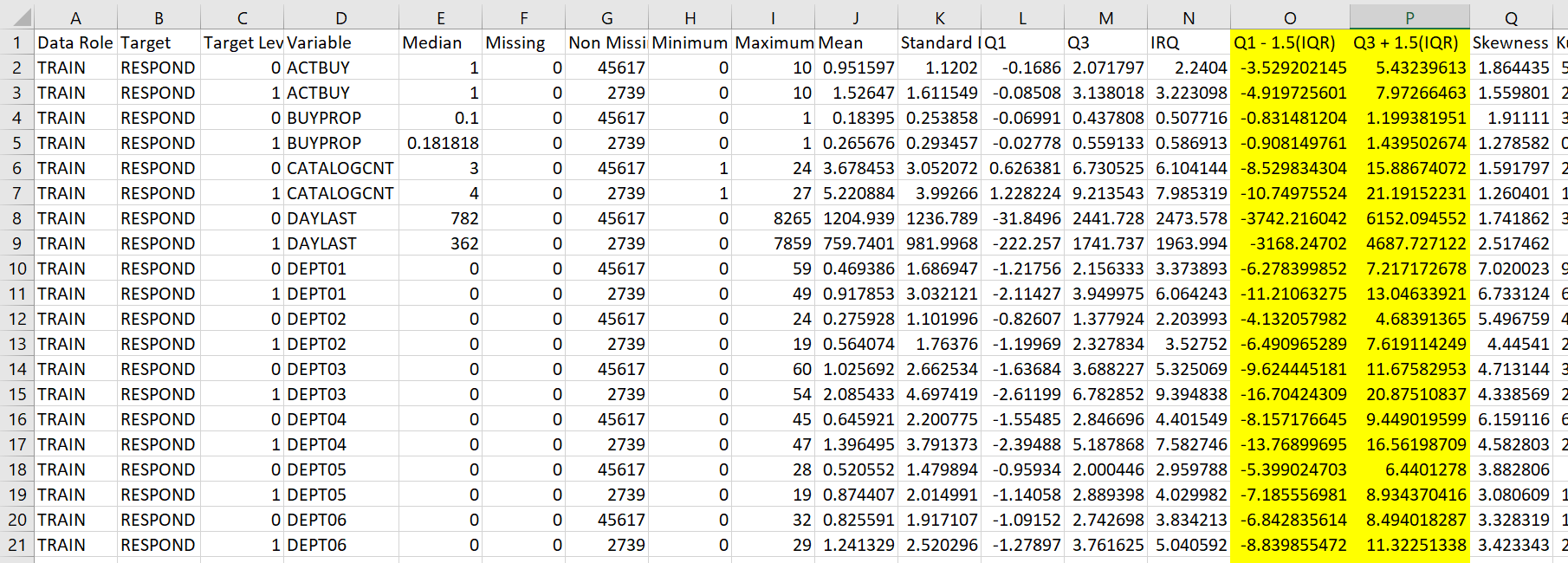
After adding and running the filter node to the project, we have the following results:



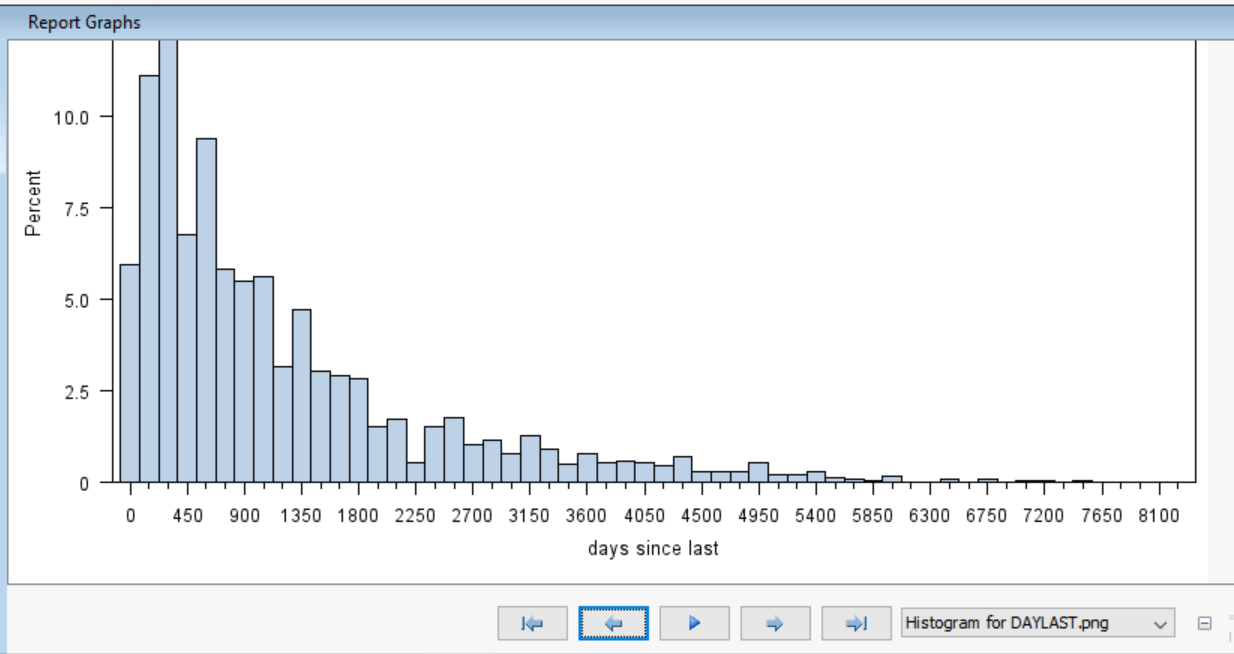
# With the StatExplore Node, we generated the Interval Variables table which displays a table of summary statistics for interval variables including mean and standard deviation.



We moved this table into Excel to calculate the Q1 & Q3, IQR and outlier values for each variable.



I compared this info to the [Univariate Histogram](https://documentation.sas.com/doc/en/emref/14.3/p1s9tet1nhuarjn1re7qw4vvfxp0.htm#n1qi0z36kzdq9kn15xgbo5w5h23p) for each variable generated from SAS Node Code.



Using the table and histograms for each variable, we were able to determine which variables have outliers that need to be removed.

IF ACTBUY > 7.97266463 then delete;

IF DAYLAST > 6152.094552 then delete;

IF DEPT05 > 8.934370416 then delete;

IF DEPT07 > 1.999186007 then delete;

IF DEPT10 > 11.19664554 then delete;

IF DEPT11 > 4.347513557 then delete;

IF DEPT15 > 6.712963144 then delete;

IF DEPT16 > 6.355999274 then delete;

IF DEPT18 > 3.775612457 then delete;

IF DEPT19 > 3.556790332 then delete;

IF DEPT20 > 1.63582488 then delete;

IF DEPT21 > 1.21526434 then delete;

IF DEPT27 > 8.91633293 then delete;

IF DOLINDEA > 199.4817483 then delete;

IF DOLLARQ01 > 104.7142525 then delete;

IF DOLLARQ04 > 112.6889834 then delete;

IF DOLLARQ05 > 119.9368083 then delete;

IF DOLLARQ06 > 108.8219948 then delete;

IF DOLNETDA > 191.7379656 then delete;

IF MONLAST > 202.0299063 then delete;

IF TOTORDQ01 > 1.847747185 then delete;

IF TOTORDQ03 > 1.373660894 then delete;

IF TOTORDQ05 > 1.958658868 then delete;

IF TOTORDQ09 > 1.813416952 then delete;

IF TOTORDQ10 > 1.86231867 then delete;

IF TOTORDQ12 > 2.48326094 then delete;

IF TOTORDQ13 > 1.985993095 then delete;

IF TOTORDQ14 > 1.784962061 then delete;

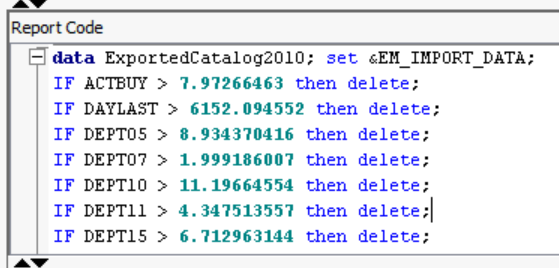
IF TOTORDQ16 > 2.658483902 then delete;

IF TOTORDQ17 > 2.261503116 then delete;

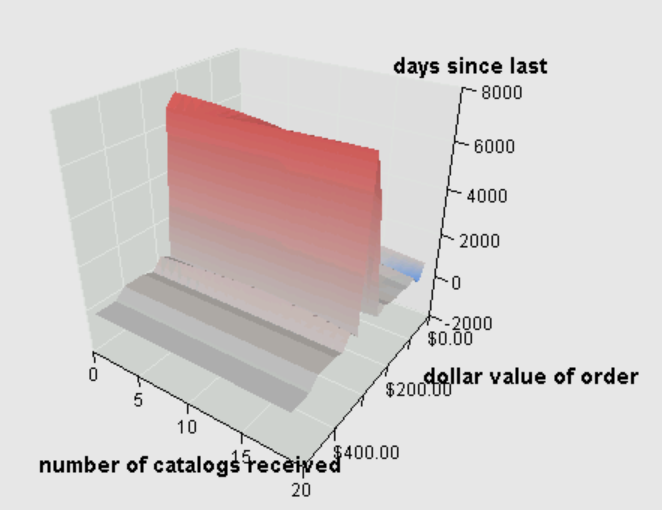
IF TOTORDQ19 > 1.964440261 then delete;

IF UNITSLAP > 105.2575375 then delete;

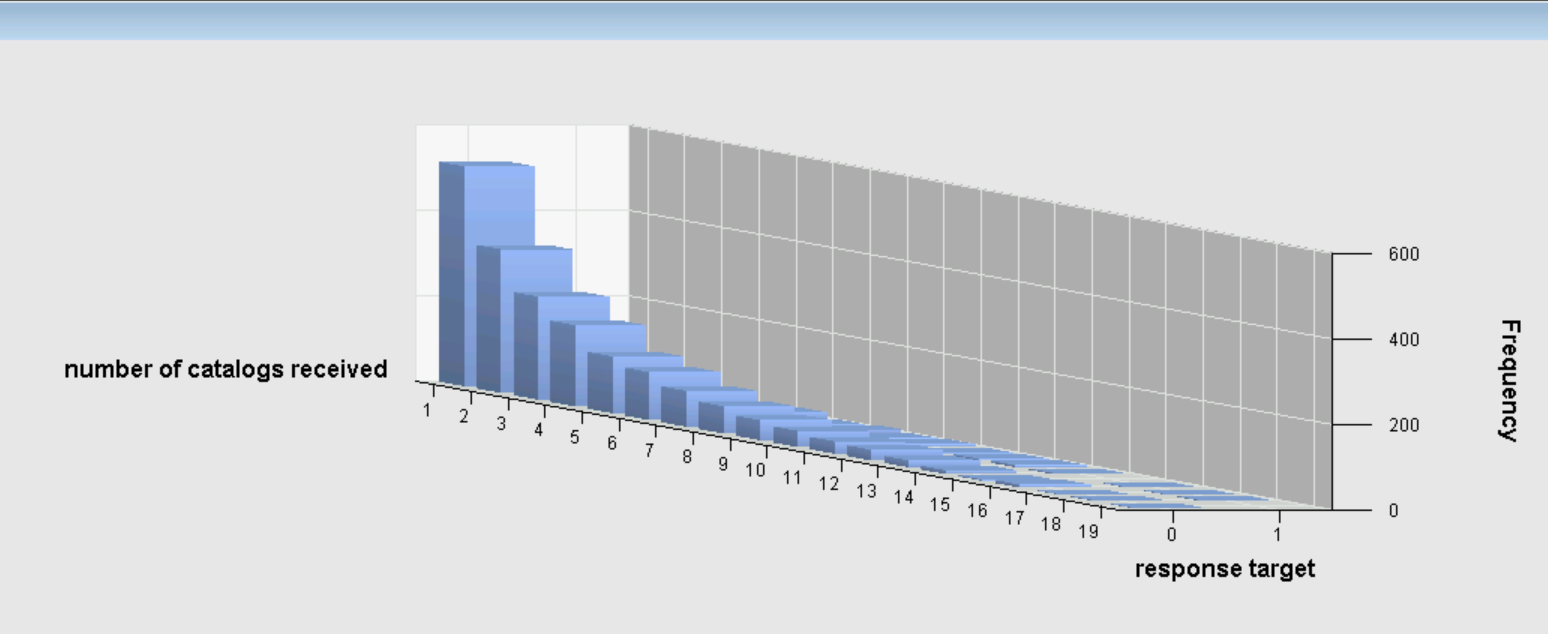
IF UNTLANPO > 12.00516903 then delete;



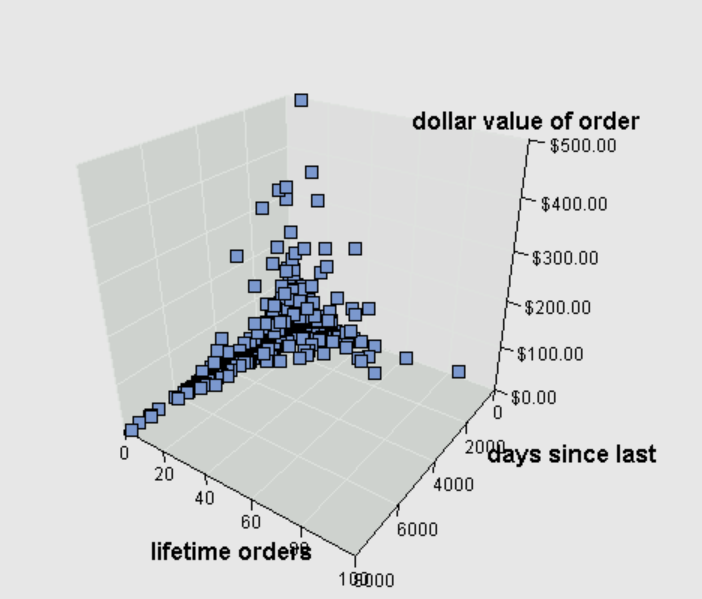
* You can make 3-D plots to get a better sense of how independent variables affect the dependent variables.



This 3D surface area graph shows the relationship between Number of catalogs received, days since last purchase and the target variable of dollar value of order. When the days since last order increases part 6000, then dollar value of the order approaches $200.00.



This bar chart with both a category and series variable indicates that most customers received 1 catalog. The number of catalogs received does not increase the response target.

 This scatter plot in 3 dimensions indicates that the dollar value of the order does not increase with the number of lifetime order or the days since last order. Customers with lifetime orders > 30 tend to have ordered in the last 2000 days. The highest dollar value orders are from customers in the lower half of lifetime number of order and the lower half of days since last order. This informs us that order dollar value does not increase with the number of lifetime orders or the number of days since last order.