### EXECUTIVE SUMMARY

Our analysis of Scholastic Travel Company data revealed 11 variables of value in helping predict which clients of the company will be repeat customers. The strength of our predictive model, based on its c statistic of 0.822, is moderately strong.

### PROBLEM STATEMENT

Our task was to build a predictive model for forecasting which Scholastic Travel Company customers are most likely to book a trip in the following year using company data from the previous years.

### LOGISTIC REGRESSION PREDICTORS

Table A

|  |  |
| --- | --- |
| **Quantitative Variables** | **Classification Variables** |
| Recode.Days  NEWVAR.Diff.Deposit.Depart.Date (Reference variable)  Recode.FRP.Active  Recoded.FRP.Cancelled  FRP.Take.up.percent  NEWVAR.Diff.Early.RPL.Depart (Reference variable)  NEWVAR.Diff.Latest.RPL.Depart (Reference variable)  RECODED.Cancelled.Pax  RECODED.Diff.Initial.Sy.Dept.Dt (Reference variable)  RECODED.L2.Total.School.Enrollmt  Recode.EZ.Pay.Take.Up.Rate  Recode.FPP  Recoded.Total.Pax  RECODED.SPR.Group.Revenue  NumerofMeetingswithParents  Recode2.DifferencTravelToFirstMe  RECODE2.DifferenceTraveltoLastMe  RECODED.L2.FPP.to.Schol.enrollmt  RECODED.FPP.to.PAX  RECODED.Num.of.Non\_FPP.PAX  NEWVAR.Ratio.CnclledPaxTotalPax | Travel.Type  Parent.Meeting.Flag  Special.Pay  School.Sponsor  SingleGradeTripFlag  Region  CRM.Segment  School.Type  Recoded.Income.Level  SPR.New.Existing  SchoolGradeType  DepartureMonth  GroupGradeType  MajorProgramCode  Recoded.SchoolSizeIndicator |

To use date variables of interest in our model, we converted the dates to numeric values by calculating the number of days between the date of interest and the departure date. The new reference variables are identified above as reference variables.

Lastly we created a new variable, NEWVAR.Ratio.CnclledPaxTotalPax, to capture Cancelled.Pax & Total.Pax information, which were identified as highly correlated variables from our correlation analysis.

### LOGISTIC REGRESSION RESULTS

The following table lists the significant variables, based on an alpha level of 0.05, from our logistic regression for Retained.in.2012 with a response of 1.

Table B

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predictor variables** | **Reference level (if any)** | ***P*-value of Chi-square statistic** |
| 1 | NEWVAR.Diff.Deposit.Depart |  | 0.0345 |
| 2 | RECODED.Diff.Initial.System |  | 0.0096 |
| 3 | RECODED.L2.FPP.to.School.Enrollment |  | 0.0175 |
| 4 | Region: Northern California |  | 0.0257 |
| 5 | Region: Pacific Northwest |  | 0.0053 |
| 6 | School.Type: Public |  | 0.0017 |
| 7 | SPR.New.Exisiting: Existing |  | < 0.0001 |
| 8 | Departure.Month : April |  | 0.0129 |
| 9 | Departure.Month: February |  | 0.0326 |
| 10 | Departure.Month: June |  | 0.0010 |
| 11 | Group.Grade.Type: K->High |  | 0.0358 |

### ANALYSIS OF REGRESSION RESULTS

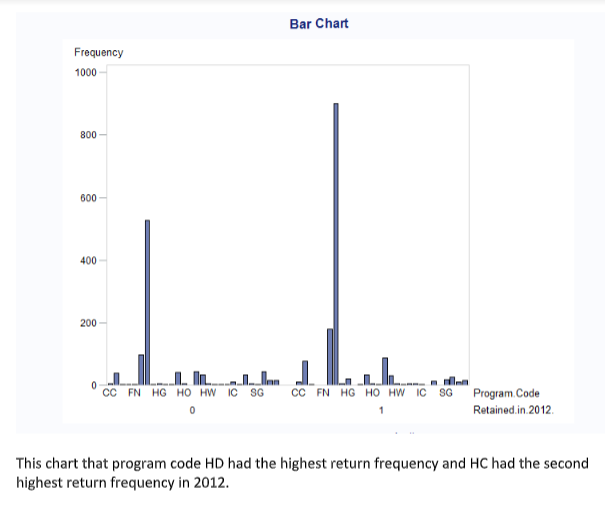
In evaluating our logistic regression model, we started by looking at the Likelihood Ratio from the Global Null Hypothesis test, which had a *p*-value below 0.0001. This is evidence that at least one of the predictor variables is not 0. Next we reviewed all predictor variables from the result of our logistic regression and identified 11 predictors with significant *p*-values (see Table B in section LOGISTIC REGRESSION RESULTS).

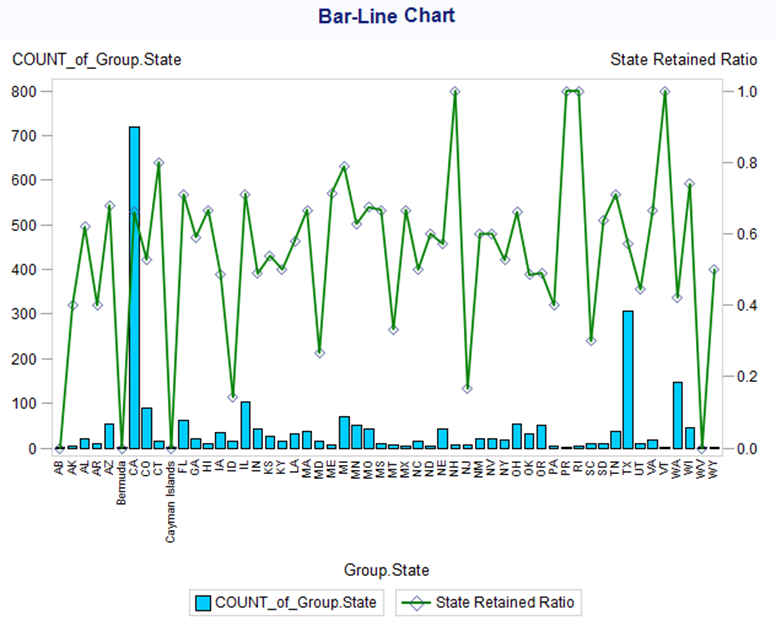
Finally a review of the rank correlation indices shows a *c* (concordance) statistic score of 0.821. With a range of possible values for the c statistic between 0.500 and 1, our score of 0.821 indicates that our final 11 predictor variables (those with significant *p*-values) are relatively strong in predicting which clients of Scholastic Travel Company will return the following year.

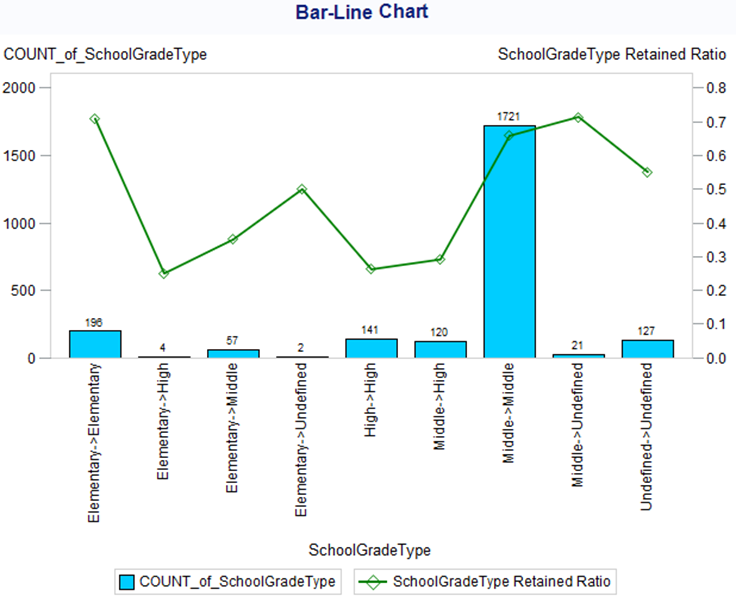
### APPENDIX

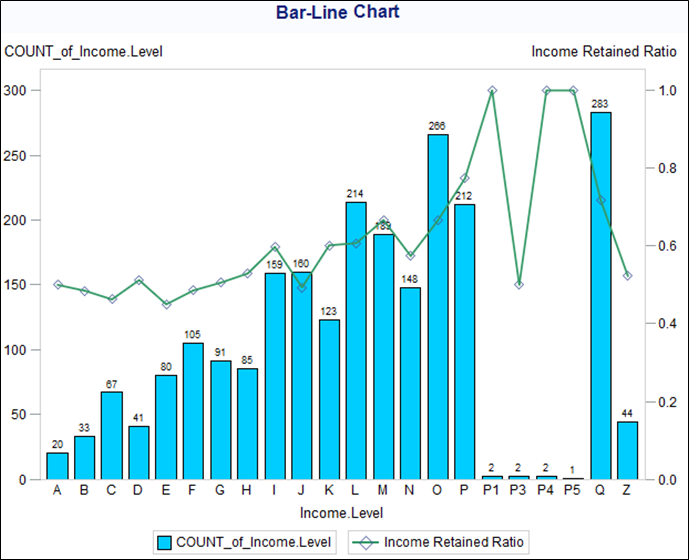
### DATA UNDERSTANDING ARTIFACTS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DATES** | **LOCATION** | **MEETING** | **PAYMENT** | **SCHOOL** | **TRIP** |
| Departure.Date | Region | Parent.Meeting.Flag | FRP.Active | From.Grade | Program.Code |
| Return.Date | Group.State | NumberOfMeetingswithParents | FRP.Cancelled | To.Grade | Days |
| Deposit.Date |  | FirstMeeting | FRP.Take.up.percent. | Is.Non.Annual. | Travel.Type |
| Early.RPL |  | LastMeeting | Cancelled.Pax | Special.Pay | Tuition |
| Latest.RPL |  | DifferenceTraveltoFirstMeeting | Total.Discount.Pax | Poverty.Code | EZ.Pay.Take.Up.Rate |
| Initial.System.Date |  | DifferenceTraveltoLastMeeting | SPR.Group.Revenue | CRM.Segment | FPP |
| DepartureMonth |  |  | School.Sponsor | School.Type | Total.Pax |
|  |  |  |  | MDR.Low.Grade | MajorProgramCode |
|  |  |  |  | MDR.High.Grade | SingleGradeTripFlag |
|  |  |  |  | Total.School.Enrollment | FPP.to.School.enrollment |
|  |  |  |  | Income.Level | FPP.to.PAX |
|  |  |  |  | SPR.New.Existing | Num.of.Non\_FPP.PAX |
|  |  |  |  | SchoolGradeTypeLow | SPR.Product.Type |
|  |  |  |  | SchoolGradeTypeHigh |  |
|  |  |  |  | SchoolGradeType |  |
|  |  |  |  | GroupGradeTypeLow |  |
|  |  |  |  | GroupGradeTypeHigh |  |
|  |  |  |  | GroupGradeType |  |
|  |  |  |  | SchoolSizeIndicator |  |







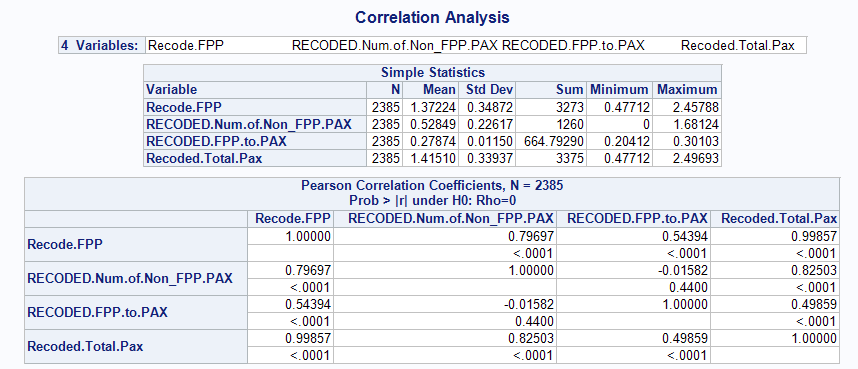


### ARTIFACTS FROM DATA PREPARATION

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Data Field Name** | **Data understanding Notes** | **Notes** |
|  | **ID** | Number of records: 2,389 |  |
| CAT | **Program.Code** | Program codes above 10% frequency: HC (11.47%), HD(59.86%). Most frequencies for programs are less than 4% with 2 above 4% | Excluded (too many levels) |
| CAT | **From.Grade** | 127 MISSING  Lowest grades in these programs with the highest frequencies are grades 6, (9.46%) 7th (21.56%) grade and 8th grade (46.92%). The rest are below 7%. 127 NA’s (Can put mean values instead of NA) | Excluded (using other grade variables) |
| CAT | **To.Grade** | 150 MISSING  68.9% of the students are in 8th grade for this column. FRom the to.grade and from.grade it seems 8th grade students are highest frequencies in both columns. 150 NA’s (Can put mean values instead of NA) | Excluded (using other grade variables) |
| CAT | **Group.State** | State with highest frequency is CA (30.05%) and TX (12.89%). Rest are below 5%. Nothing seems off about this column and is ready to be analyzed | Excluded (too many levels) |
| CAT | **Is.Non.Annual.** | Binary data set with no missing values. 84.6% 0s | Excluded (Highly correlated with response variable) |
| NUM | **Days** | No missing values, Has outliers positvely skewed. Average number of days the groups are on the program is 4.57 days. 50th persentile is 5 days. | Positive skew fix (done) |
| CAT | **Travel.Type** | Travel.type A dominates this column with a frewuency of 84.30%. Most people like air travel.nothing wrong with this dataset, this set is types of travel and is in letter format. each letter has meaning for the type | No missing values |
| CAT | **Departure.Date** | June through July, there is a surge in freqeuncy. no missing data, no outliers | Excluded (use Days variable) |
| CAT | **Return.Date** | No outliers, Surge in demand in june to july. no missing data, no outliers | Excluded (use Days variable) |
| CAT | **Deposit.Date** | NO MISSING.  Some date values with very high freq (date/%):  15OCT2010 16.16; 01NOV2010 10.30; 01OCT2010 9.96; 29OCT2010 6.82; 22OCT2010 6.53; 15NOV2010 5.73; 19NOV2010 4.90  Other than above, most dates have proportions < 1% and a few others with proportions b/w 1% and 3% | Use NEWVAR.Diff.Deposit.Depart.Date |
| CAT | **Special.Pay** | 2 missing observations.  Elements of variable: CP, FR, NA, SA.  Frequency appears normally distributed  NOTE: NA here means normal arrangement. | Filtered out missing values using filter pane in query builder |
| NUM | **Tuition** | Data is skewed negatively. Moerate number of outliers.  NOTE: in analysis account for unique high cost of West-cost air trips | Negative skew fix (done), No missing values |
| NUM | **FRP.Active** | Many outliers. Distribution skewed right.  There were trips with 0 insurance. | No missing (positive skew) |
| NUM | **FRP.Cancelled** | Many outliers. Distribution skewed right. | positive skew (no missing) |
| NUM | **FRP.Take.up.percent.** | This is ration of FRP.Active/FPP.  Normally distributed with medium # of outliers | No transformation needed |
| CAT | **Early.RPL** | 673 missing observations.  Distribution appears bimodal.  NOTE: could correlate well w outcome variable | Use NEWVAR.Diff.Early.RPL.Depart:  IF not missing THEN DATDIF(t1.'Early.RPL'n, t1.'Departure.Date'n, 'Actual')  ELSE 0 |
| CAT | **Latest.RPL** | 19 missing observations. Normally distributed.  NOTE: missing for Latest.RPL not match missing in Early.RPL | Use NEWVAR.Diff.Latest.RPL.Depart:  IF not missing THEN DATDIF(t1.'Latest.RPL'n, t1.'Departure.Date'n, 'Actual')  ELSE 0 |
| NUM | **Cancelled.Pax** | Distribution skewed left.  Moderate # of outliers. | RECODED as RECODED.Cancelled.Pax (log 10)  Positively skewed with no missing values |
| NUM | **Total.Discount.Pax** | Data skewed right. Medium # of outliers | RECODED as RECODED.Total.Discount.Pax (log 10)  Positively skewed with no missing values |
| CAT | **Initial.System.Date** | 8 missing obs. Three spikes in frequency distribution.  NOTE: might have high correlation with response variable. Useful to group dates by semester? | Filtered out missing values using filter pane in query builder  RECODED as NEWVAR.Diff.Initial.Sy.Dept.Dt |
| CAT | **Poverty.Code** | 599 missing obs. Normally distributed  Elements of variable: [0, A, B, C, D, E]. B code is 53.69%.  NOTE: missing probably "Space if DISTCLASS = U (Supervisory Union)" from description | excluded too many missing (will use Income.Level) |
| CAT | **Region** | Normally distributed. No missing  Region elements [Dallas, Houston, Northern California, Other, Pacific Northwest, Southern California]  NOTE: useful to create more regions? different aggregations? | No transformation needed |
| CAT | **CRM.Segment** | 4 missing (NA)  High frequencies: 5 32.99%, 10 38.26% | Filtered out missing values using filter pane in query builder |
| CAT | **School.Type** | No missing  Elements [CHD, Catholoic, PUBLIC, Private non-Christian] Public is 76.10% | No transformation needed |
| CAT | **Parent.Meeting.Flag** | No missing  Binary data. Data related to meetings with parents. | No transformation needed |
| CAT | **MDR.Low.Grade** | 68 missing obs. Bimodal freq distribution. Data range is [PK to 10th grade]  High freq: 6 38.26%, K 18.44%; PK 17.28%, 7 14.99% | excluded (will use other grade columns) |
| CAT | **MDR.High.Grade** | 68 missing obs. Bimodal distribution. Data range [1 to 12]  High freq: 8 69.44%; 12 14.99% | excluded (will use other grade columns) |
| NUM | **Total.School.Enrollment** | 91 missing. Skewed right. Many outliers  Mean: 648; Mode: 700; Median: 597 | RECODED (replaced missing with mean and performed log 10)  **USE NEW COLUMN: RECODED.L2.Total.School.Enrollmt** |
| CAT | **Income.Level** | Alternative to poverty code. High cardinality (A-P, P1, P3, P4, P5, Q, Z) P1, P3, P4, P5 have low counts of 1 or 2. Could be typos?  62 missing.  Potentially use this variable to fill in missing values for poverty code. | RECODED P1,P3,P4,P5 as P  RECODED missing as MMMM |
| NUM | **EZ.Pay.Take.Up.Rate** | Skewed right. Medium outliers  Mean: 0.208; Mode: 0; Median: 0.20. | RECODED (log transform) |
| CAT | **School.Sponsor** | Binary variable : 0 (89.41%); 1 (10.59%). | No transformation needed |
| CAT | **SPR.Product.Type** | Variable elements: CA History (7.33), Costa Rica (1.93), East Coast (83.93), IL History (0.21), International (0.63), Science (5.99) | No transformation needed |
| CAT | **SPR.New.Existing** | Binary variable.  VAluees: New (32.73); Existing (67.27) | No transformation needed |
| NUM | **FPP** | Many outliers. Skewed right.  Mean: 31.297; Mode: 8; Median: 23 | RECODED (log transform) |
| NUM | **Total.Pax** | Many outliers. Skewed right.  Mean: 34.251; Mode: 8.0; Median: 26.0 | RECODED (log transform) |
| NUM | **SPR.Group.Revenue** | Few outliers. Skewed left.  Mean: 1615.22; Mode: 1999.0; Median: 1700.0  Maximum of 4,200 seems kinda low for such service | RECODED (log transform) |
| CAT | **NumberOfMeetingswithParents** | Count of number of meetings for specific trip. Range values: 0 to 2 (0: 337; 1: 1,471; 2: 581) | No transformation needed |
| CAT | **FirstMeeting** | 337 trips had no meetings (337 missing).  SAS data type is char -> should be date | EXCLUDED (will use differencetravel columns) |
| CAT | **LastMeeting** | 337 trips had no meetings (337 missing).  SAS data type is char -> should be date | EXCLUDED (will use differencetravel columns) |
| NUM | **DifferenceTraveltoFirstMeeting** | 337 trips had no meetings (337 missing)..  In SAS as char (need numberic)  Two unexpected negative values (-204,-188)  Lots of outliers.  Distribution is skewed right | RECODED imputed mean to missing  RECODED negatives to positive values  Parent.Meeting.Flag will capture trips with no meeting |
| NUM | **DifferenceTraveltoLastMeeting** | 337 trips had no meetings (337 missing).  In SAS as char (need numberic)  4 unexpected negative values (-17,-188,-204,-4).  Lots of outliers  Normally distributed | RECODED imputed mean to missing  RECODED negatives to positive values  Parent.Meeting.Flag will capture trips with no meeting |
| CAT | **SchoolGradeTypeLow** | Elements: Elementary, HIgh, Middle, Undefined.  Middle 77.94%; Undefined: 5.32%  'Undefined' with 127 values | EXCLUDED (will use SchoolGradeType columns) |
| CAT | **SchoolGradeTypeHigh** | Elements: Elementary, Middle, High, Undefined  Middle 74.42%; Undefined 6.28%  'Undefined' (150 obs) considered a level | EXCLUDED (will use SchoolGradeType columns) |
| CAT | **SchoolGradeType** | Middle->Middle school category is 72.04% of data  4 categories have very low counts (3% and less)  Undefined considered a level  Low freq.: Elementary->High(4); Elementary->Undefined(2) | No transformation needed |
| CAT | **DepartureMonth** | Elements: January 0.38, February 2.05, March 16.20, April 22.35, May 21.22, June 37.80 | No transformation needed |
| CAT | **GroupGradeTypeLow** | Uniform distribution. No missing values. No outliers. However, there is a group called  'Undefined' with 68 values | EXCLUDED (will use GroupGradeType) |
| CAT | **GroupGradeTypeHigh** | Skewed right. Undefined school group which consists of 2.85% of data.  'Undefined' with 68 values | EXCLUDED (will use GroupGradeType) |
| CAT | **GroupGradeType** | Low frequency in GroupGradeType= Elementary-elementary AND Elementary- High. Non uniform distribution  Low freq: Elementary->Elementary (7); Elementary->High (8) | No transformation needed |
| CAT | **MajorProgramCode** | Non uniform distribution. No missing values. Low frequency(0.63%) in Code I (16 values). | No transformation needed |
| CAT | **SingleGradeTripFlag** | No uniform distribution because data is 1 and 0. No missing values | No transformation needed |
| NUM | **FPP.to.School.enrollment** | Skewed right. Plenty of outliers to the right.  91 missing | RECODED missing to column mean and log 10  **USE NEW COLUMN: RECODED.L2.FPP.to.Schl.enrollmt**  Also, same number of missing (91) as FPP.to.School.enrollment |
| NUM | **FPP.to.PAX** | Uniform distribution. Outliers present when values are less than 0.65. | RECODED (log transform) |
| NUM | **Num.of.Non\_FPP.PAX** | Heavily skewed right. Plenty of outliers when number exceeds 21.3. | RECODED (log transform) |
| CAT | **SchoolSizeIndicator** |  | RECODED missing values to MMMM  NOTE: Also, same number of missing (91) as FPP.to.School.enrollment |
|  | Retained.in.2012. | 0 = 39.26%  1 = 60.74% |  |

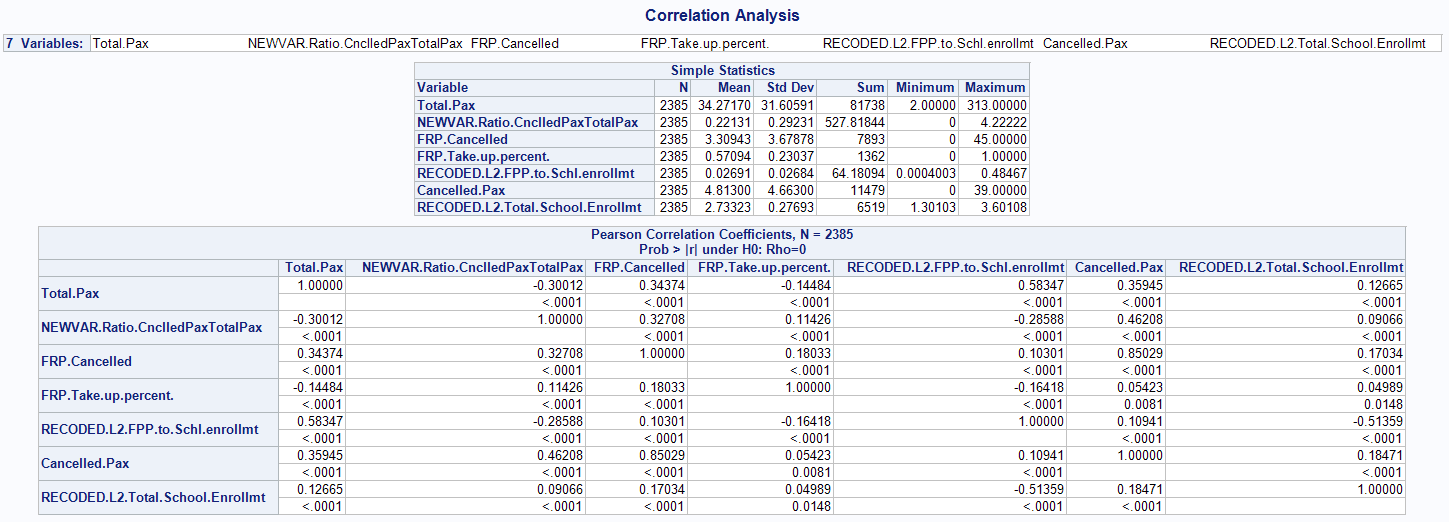
**MODELING ARTIFACTS**

Correlation test of variables related to FPP and PAX



* Keeping Num.of.Non\_FPP.PAX and FPP.to.PAX (both ratios) from variables related to passenger counts. These are the only variables NOT correlated to one another.

Correlation test of variables related to FRP and Enrollment



Based on above correlation analysis, we kept the following variables:

* FPP.to.School.Enrollment (ration) (don't keep Total.Pax)
* Cancelled.Pax
* FRP.Cancelled

We created a reference variable, NEWVAR.Ratio.CnclledPaxTotalPax, to capture Cancelled.Pax & Total.Pax information, which were highly correlated.