Paper 104-2019

Using Factor Analysis and Multivariate Analysis of Variance to Explore Academic Achievement in the 2016 Monitoring the Future Study

Stefanie N. Kairs, National University, San Diego, CA

ABSTRACT

The 2016 Monitoring the Future survey is part of an annual, long-term study of American adolescents and adult high school graduates conducted by the University of Michigan's Institute for Social Research. This secondary data analysis uses the FACTOR procedure in the SAS^{TM} Studio software to perform factor analysis to extract latent structures describing academic achievement, environment, and student delinquency. 17,719 observations were used to perform multivariate analysis of variance via the GLM procedure to explore the relationships between the extracted factors and demographic variables for ethnicity, gender, and population density. The SAS Code and results are presented, along with a discussion of the necessary data cleaning steps, data quality assessment and post-hoc analyses. A statistically significant overall effect was found for each independent variable: gender (Wilks' Lambda = 0.96, F= 273.23, df= (3, 17711), p < 0.0001); race (Wilks' Lambda = 0.86, F= 479.83, df= (6, 35422), p < 0.0001); population density (Wilks' Lambda = 0.98, F= 65.31, df= (6, 35422), p < 0.0001). Population density explains 2% (Pillai's trace = 0.022, p < 0.0001) of the variance in academic achievement, academic environment, and at-risk behaviors. Gender explains 4% (Pillai's trace = 0.044, p < 0.0001), and race explains 14% (Pillai's trace = 0.145, p < 0.0001) of the variance. The academic environment for 8th and 10th grade students was described by an extracted factor with high loadings for the variables for parental education, college preparatory program, and remedial schooling (negative loading), and was shown to vary significantly by race.

INTRODUCTION

Previous research has shown that urbanicity is an important indicator for understanding adolescent academic achievement (Miller & Vortruba-Drzal, 2015) and that gender differences in achievement favoring female students persist across subject domains (Voyer, Voyer, & Hinshaw, 2014). Additionally, gender differences remain significant after controlling for socio-economic background (Carvalho, 2016). This study uses SAS^{TM}[1] Studio software and a publicly available dataset to explore the relationships between demographic variables for race, gender, and population density and extracted factors describing academic achievement, academic environment and at-risk behaviors. The FACTOR procedure will be used to perform factor analysis and extraction. The GLM procedure will be used to perform multivariate analysis of variance (MANOVA).

The 2016 Monitoring the Future (MTF) 8th- and 10th-Grade surveys are part of an annual, long-term study of American adolescents and adult high school graduates conducted by the University of Michigan's Institute for Social Research. Study data have been collected since 1975, when the Institute of Social Research opened. The 2016 MTF survey involved 45,500 participants in grades 8, 10 and 12, and sampled from 372 secondary schools across the country (Johnston, O'Malley, Miech, Bachman, & Schulenberg, 2017). With an estimated 500 variables per year, the survey covers a broad range of topics including drug use, attitudes and beliefs regarding drug use, and lifestyle choices and values (Johnston, Bachman, O'Malley, Schulenberg, & Miech, Codebook., 2016). The data and codebook are publicly available via the Inter-University Consortium for Political and Social Research (ICPSR) website (Johnston, Bachman, O'Malley, Schulenberg, & Miech, 2017). Factor analysis can be used to mathematically define underlying data structures which reflect characteristics that cannot or were not directly queried. This secondary data analysis will use those techniques to extract the factors for student achievement, their academic attitudes and environment, and at-risk behaviors or delinquency. The extracted factors will then be used as the dependent variables in multivariate analysis of variance (MANOVA). The MANOVA analysis will evaluate the relationship between the demographic independent

[1] Copyright © 2012-2017, SAS Institute Inc., Cary, NC, USA. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA

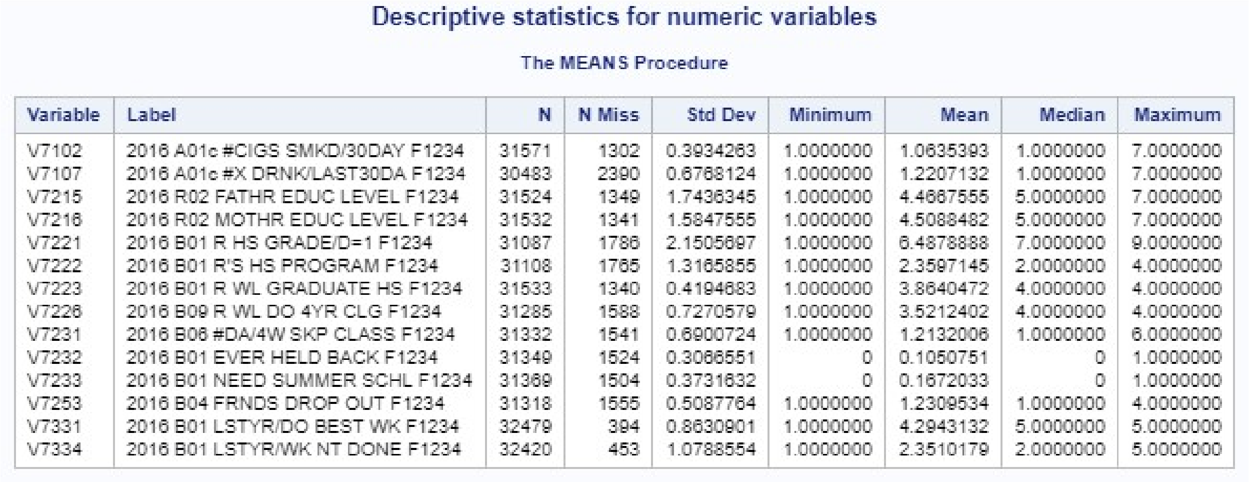
variables for gender, ethnicity, and population density and the dependent factors for academic success, academic environment, and at-risk behaviors. Race, gender, and urbanicity are hypothesized to have significant relationships with academic achievement and delinquency, and gender is not expected to be statistically significantly related to academic environment.

DATA QUALITY

The 2016 MTF dataset for 8th- and 10th-Grade survey responses has 32,873 responses for 571 variables. Four survey forms were used for data collection at the 8th and 10th grade levels and not all questions were asked on all forms (Johnston, Bachman, O'Malley, Schulenberg, & Miech, Codebook., 2016). Prior exploratory analysis of the 2016 MTF dataset indicated that underlying factors could be identified and extracted to represent student academic achievement, attitudes toward education, and delinquent or at-risk behaviors. A subset of 14 variables from the exploratory analysis were chosen for factor analysis and were hypothesized to represent these three major groupings. To increase the number of complete cases for factor analysis, the selected variables were chosen from those that appeared on all four survey forms, where possible. Demographic variables for grade level, gender, ethnicity, and metropolitan statistical area (MSA) were retained to facilitate the MANOVA analysis and investigations into missing data. Analysis was performed in SAS^{TM} Studio, release 3.7 (Enterprise Edition, Build Jun 11, 2018). The supporting code is available in Appendix A.

DATA QUALITY EVALUATION

The MEANS procedure was used to investigate missing data and extreme values in the 14 variables selected for factor analysis. The results are presented in Table 1. Less than 7.5% of the data were missing for each variable. Variables except respondent's high school program type (V7222) were either dichotomous (0-1, or 1-2), rated on a Likert scale (1-4, or 1-5), ordinal, or discretized continuous variables represented on an ascending scale. The variable for high school program type consisted of four nominal levels for high school program with no implicit ranking beyond "college prep" vs "all others." No variables had scores out of range or extremely low variance. The variables for high school program type (V7222), father's education level (V7215) and mother's education level (V7216) will be re-coded in the following sections to ease interpretation of the factor analysis.



Descriptive statistics for numeric variables

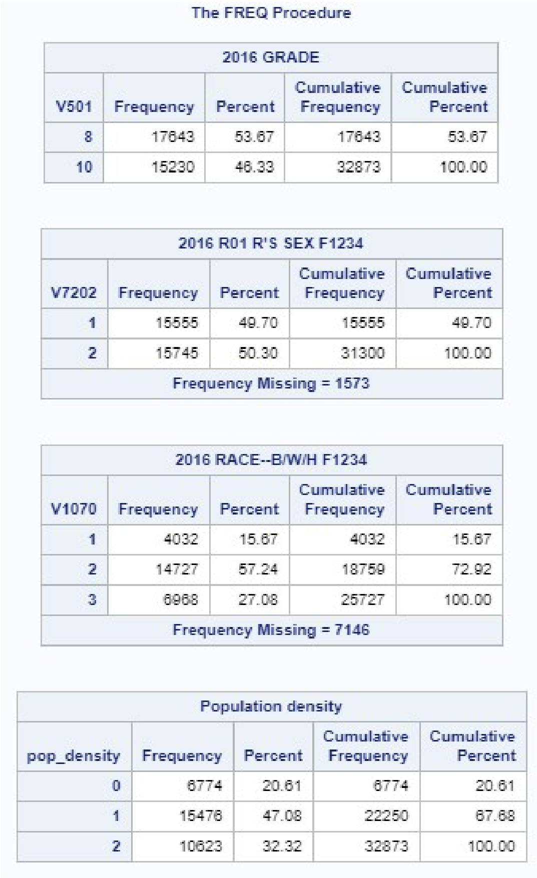
The MEANS Procedure

Variable	Label	N	N Miss	Std Dev	Minimum	Mean	Median	Maximum
V7102	2016 A01c #CIGS SMKD/30DAY F1234	31571	1302	0.3934263	1.0000000	1.0635393	1.0000000	7.0000000
V7107	2016 A01c #X DRNK/LAST30DA F1234	30483	2390	0.6768124	1.0000000	1.2207132	1.0000000	7.0000000
V7215	2016 R02 FATHR EDUC LEVEL F1234	31524	1349	1.7436345	1.0000000	4.4667555	5.0000000	7.0000000
V7216	2016 R02 MOTHR EDUC LEVEL F1234	31532	1341	1.5847555	1.0000000	4.5088482	5.0000000	7.0000000
V7221	2016 B01 R HS GRADE/D=1 F1234	31087	1786	2.1505697	1.0000000	6.4878888	7.0000000	9.0000000
V7222	2016 B01 R'S HS PROGRAM F1234	31108	1765	1.3165855	1.0000000	2.3597145	2.0000000	4.0000000
V7223	2016 B01 R WL GRADUATE HS F1234	31533	1340	0.4194683	1.0000000	3.8640472	4.0000000	4.0000000
V7226	2016 B09 R WL DO 4YR CLG F1234	31285	1588	0.7270579	1.0000000	3.5212402	4.0000000	4.0000000
V7231	2016 B06 #DA/4W SKP CLASS F1234	31332	1541	0.6900724	1.0000000	1.2132006	1.0000000	6.0000000
V7232	2016 B01 EVER HELD BACK F1234	31349	1524	0.3066551	0	0.1050751	0	1.0000000
V7233	2016 B01 NEED SUMMER SCHL F1234	31369	1504	0.3731632	0	0.1672033	0	1.0000000
V7253	2016 B04 FRNDS DROP OUT F1234	31318	1555	0.5087764	1.0000000	1.2309534	1.0000000	4.0000000
V7331	2016 B01 LSTYR/DO BEST WK F1234	32479	394	0.8630901	1.0000000	4.2943132	5.0000000	5.0000000
V7334	2016 B01 LSTYR/WK NT DONE F1234	32420	453	1.0788554	1.0000000	2.3510179	2.0000000	5.0000000

Table 1. Descriptive Statistics for Numeric Variables before Recoding or Parceling

Variables V508 and V509 are component variables that together specify a standardized 3-category measure of population density, indicating whether the associated region is a large or medium Metropolitan Statistical Area (MSA) or fails to qualify for the MSA designation and has been scored "non-MSA." These variables were additively combined to produce a measure for population density, "pop_density," as described in the 2016 MTF Codebook (Johnston, Bachman, O'Malley, Schulenberg, & Miech, Codebook., 2016). The SUM function in SAS was not used, as assigning a zero value to missing data for either variable would yield a misleading population density designation.

The '+' operator was used so that a missing value for either V508 or V509 would result in a missing value for pop_density. The FREQ procedure was used to investigate missing data and proportions of the categorical demographic variables retained for MANOVA analysis and the student grade level (V501). The proportions of male and female respondents were relatively similar. 22% of observations were missing data for ethnicity. Over half (57.2%) of the respondents who reported their race were White (V1070=2). Almost half (47.1%) of respondents were from moderately-populated areas (pop_density=1).



The FREQ Procedure

2016 GRADE				
V501	Frequency	Percent	Cumulative Frequency	Cumulative Percent
8	17643	53.67	17643	53.67
10	15230	46.33	32873	100.00

2016 R01 R'S SEX F1234				
V7202	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	15555	49.70	15555	49.70
2	15745	50.30	31300	100.00
Frequency Missing = 1573				

2016 RACE--B/W/H F1234				
V1070	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	4032	15.67	4032	15.67
2	14727	57.24	18759	72.92
3	6968	27.08	25727	100.00
Frequency Missing = 7146				

Population density				
pop_density	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6774	20.61	6774	20.61
1	15476	47.08	22250	67.68
2	10623	32.32	32873	100.00

Figure 1. Frequencies and Proportions for Categorical Demographic Variables

The UNIVARIATE procedure was used to examine the distributions of the numeric variables. A highly right-skewed distribution was observed for V7102 (CS = 9.08), such that the 95th percentile of all observations was 1, the lowest value in a 7-point scale for the number of cigarettes smoked in the past 30 days. A response of 1 corresponds with no cigarette use at all. This right-skewed distribution is consistent with findings of declining smoking initiation and cigarette use in adolescents by the MTF principal investigators (Johnston, O'Malley, Miech, Bachman, & Schulenberg, 2017).

DATA CLEANING AND PARCEL CREATION

Variables for high school program type (V7222), father's education level (V7215), mother's education level (V7216), repeating a grade (V7232), and summer school (V7233) were recoded and/or parceled prior to factor analysis. Variable V7222 consisted of four nominal levels for high school program and was recoded as a dichotomous variable for whether the student was in a college preparatory program (0=no, 1=yes). For parental education levels, V7215 and V7216, the "don't know" category was originally coded as level 7 of an ordered scale of ascending levels of education (1= grade school, 6= grad school). These values for "don't know" were recoded as missing, and V7215 and V2716 were additively combined to a parcel for parent education (parent_ed). A parcel for remedial schooling (rem_school) additively combines the dichotomous variables for whether a respondent has ever needed summer school (V7233) or had ever been held back a grade (V7232). V7232 and V7233 showed low variance in the initial data quality assessment (Table 1). These recoding and parceling steps are consistent with the prior analysis that identified the underlying factors of interest.

MISSING DATA

To assess the impact of missing data on the number of complete cases for analysis, a dummy variable “anymiss” was created which was coded 1 when the respondent had one or greater missing values across all variables and 0 when the case was complete across the newly created parcels, the numeric variables for factor analysis, and the population density demographic variable. 17,719 (53.9%) observations were complete for these 15 variables. A second dummy variable, “missdata,” was created to reflect the count of the number of missing values for each record. 95% of records were missing values for 4 or fewer variables. The PROC FREQ output for these variables is presented in Appendix B, Figure 7. Cross-tabulation tables of anymiss and grade level, gender, ethnicity were populated, and Pearson’s chi-square test for association was performed for each comparison. These tables and the output of the statistical tests are presented in Appendix B, Figure 8 - Figure 10. Statistically significant associations were found between anymiss and all demographic variables investigated, while the proportion of complete cases was within ± <5% of the observed proportion of complete cases in the entire sample, for all comparisons except ethnicity. A t-test was performed to evaluate the differences in the number of missing values (missdata) by grade level (V501). Though the t-test (unequal variances) demonstrates a statistically significant difference in means for grade 8 and grade 10 (t = 13.30, df = 32,862, p <0.0001), the observed difference in means is less than one missing item (0.2842) and is not practically significant.

Figure 2 shows the histograms for missdata by grade level. Both distributions are highly right-skewed and span a similar range of responses. See Appendix B, Figure 11 for the TTEST procedure’s statistical output.

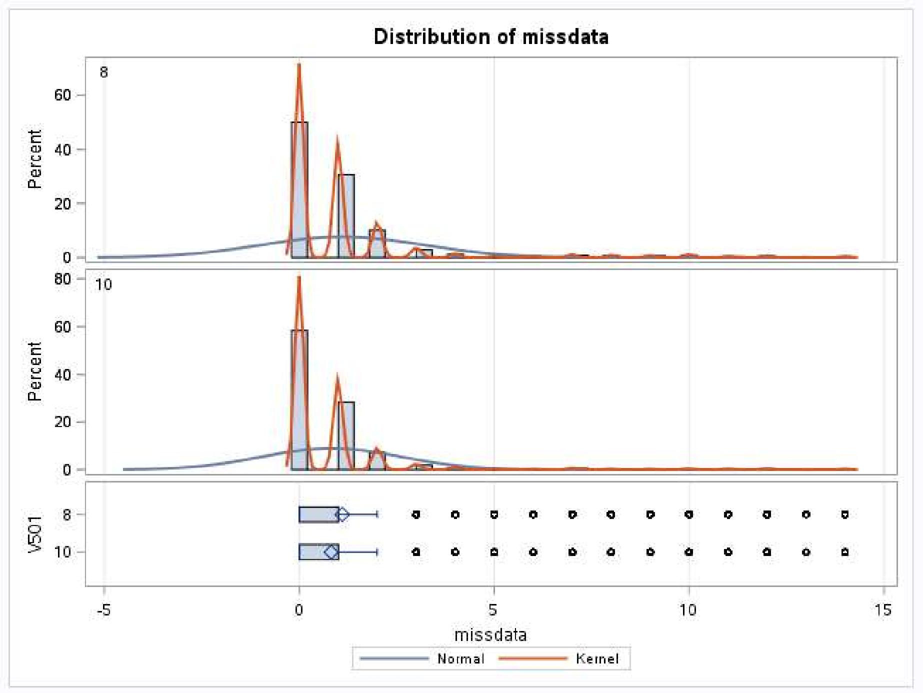


Figure 2. Distributions of Count of Missing Values by Grade

FACTOR ANALYSIS

Factor analysis was performed in SAS^{TM} Studio using the FACTOR procedure. The REORDER option was used to sort variables by their factor loadings and the SCREE option was used to produce a scree plot. Computation of the parallel analysis criterion for factor retention was performed using a script previously published by Brian O’Connor (2000). The parallel analysis criterion offers an alternative to the

Kaiser criterion, which may retain too many factors, and visual inspection of the scree plot, which may be considered too subjective. The eigenvalue of a factor should exceed the 95th percentile eigenvalue of the similarly-numbered factor as generated by the parallel analysis criterion script. The output from this script is presented in Figure 12, Appendix B. Principal components analysis with a communality estimate of one was used for extracting factors. The Kaiser criterion, parallel analysis criterion, and scree plot shown in Figure 3 each suggest three factors, as expected. For 22,079 observations, the 95th percentile of eigenvalues for the fourth factor from randomly generated data was 1.020. The eigenvalue for factor 4 in this analysis was 0.977. The table of Eigenvalues is given in Figure 13, Appendix B.

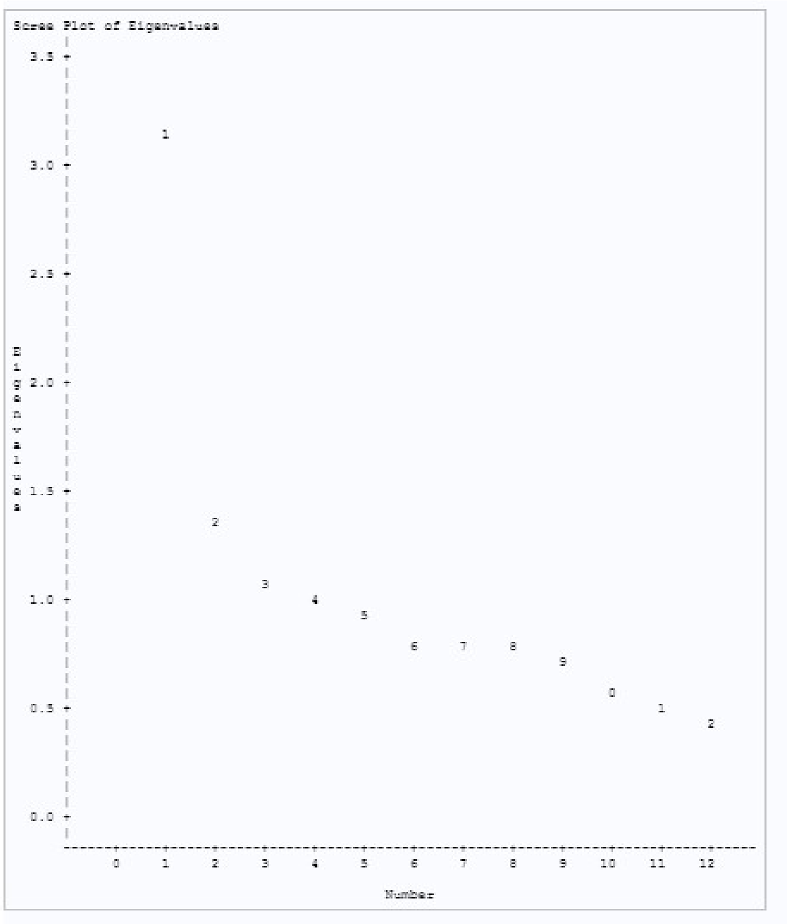
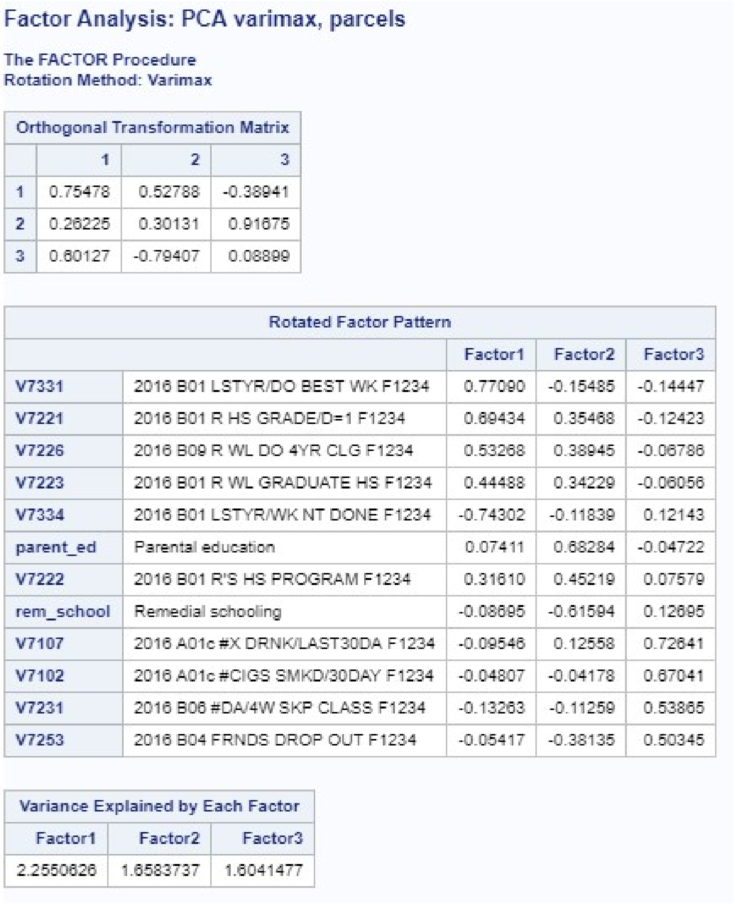


Figure 3. Scree Plot from Factor Analysis

Figure 4 shows the resulting rotated factor pattern, after varimax rotation. The first factor, representing academic achievement and attitudes, had high loadings (>0.4) for “I often do my best work,” high school grades, and the student’s projections for whether he or she would graduate and whether he or she would attend a 4-year college. This factor loaded negatively on “I often fail to complete or turn in my assignments.” The second factor, representing academic environment, had high loadings for parental education and college preparatory program, and negative loading for remedial schooling interventions. The last factor, representing delinquent or at-risk behaviors, had high loadings for drinking, alcohol use, friends who had dropped out, and skipping class.



Factor Analysis: PCA varimax, parcels

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix			
	1	2	3
1	0.75478	0.52788	-0.38941
2	0.26225	0.30131	0.91675
3	0.60127	-0.79407	0.08899

Rotated Factor Pattern				
		Factor1	Factor2	Factor3
V7331	2016 B01 LSTYR/DO BEST WK F1234	0.77090	-0.15485	-0.14447
V7221	2016 B01 R HS GRADE/D=1 F1234	0.69434	0.35468	-0.12423
V7226	2016 B09 R WL DO 4YR CLG F1234	0.53268	0.38945	-0.06786
V7223	2016 B01 R WL GRADUATE HS F1234	0.44488	0.34229	-0.06056
V7334	2016 B01 LSTYR/WK NT DONE F1234	-0.74302	-0.11839	0.12143
parent_ed	Parental education	0.07411	0.68284	-0.04722
V7222	2016 B01 R'S HS PROGRAM F1234	0.31610	0.45219	0.07579
rem_school	Remedial schooling	-0.08695	-0.61594	0.12695
V7107	2016 A01c #X DRNK/LAST30DA F1234	-0.09546	0.12558	0.72641
V7102	2016 A01c #CIGS SMKD/30DAY F1234	-0.04807	-0.04178	0.67041
V7231	2016 B06 #DA/4W SKP CLASS F1234	-0.13263	-0.11259	0.53865
V7253	2016 B04 FRNDS DROP OUT F1234	-0.05417	-0.38135	0.50345

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.2550626	1.6583737	1.6041477

Figure 4. Rotated Factor Pattern (Varimax)

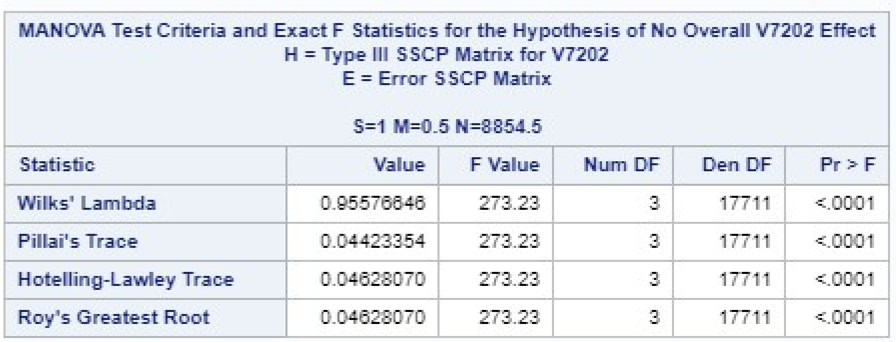
MANOVA

Multivariate Analysis of Variance (MANOVA) was performed to explore the relationship between each of the independent variables, gender, race, and population density, and the dependent variables, our extracted factors for academic success (Factor 1), academic environment (Factor 2), and delinquency or at-risk behaviors (Factor 3). The null hypotheses are: There is no relationship between gender and Factors 1-3; There is no relationship between ethnicity and Factors 1-3; and there is no relationship between population density and Factors 1-3.

MANOVA was performed in SAS^{TM} Studio using the GLM procedure (generalized linear model). The PLOTS = ALL option was used to request all applicable plots, but the output was restricted to plots with less than 5,000 points by the Output Delivery System (ODS) settings. The independent variables were the categorical variables for gender (V7202), race (V1070) and population density (pop_density). Gender has two levels, male (1) and female (2). Race has three levels, Black (1), White (2), and Hispanic (3). Population density has three levels, low-density (0), moderate-density (1), and high-density (2). For population density, we are primarily interested in the differences between high-density population centers and all other environments, so a contrast statement will be used to code this comparison. All hypotheses will be evaluated in the initial analysis. If necessary, post-hoc analysis will be performed by Tukey's Studentized Range Test using a Bonferroni correction for multiple testing.

RESULTS

Observations for 17,719 (53.9%) respondents were used to perform multivariate analysis of variance in SAS^{TM} Studio using PROC GLM. The dependent variables were factors for academic achievement, academic environment, and delinquency/at-risk behavior. A statistically significant overall effect was found for each independent variable: gender (Wilks' Lambda = 0.96, F= 273.23, df= (3, 17711), p < 0.0001); race (Wilks' Lambda = 0.86, F= 479.83, df= (6, 35422), p < 0.0001); population density (Wilks' Lambda = 0.98, F= 65.31, df= (6, 35422), p < 0.0001). The test statistics table for gender is shown in Figure 5. The tables for race and population density are given in Figure 14 and Figure 15, Appendix B. The null hypothesis of no overall effect for each independent variable was rejected. For population density, the contrast of high-density vs. all others was also statistically significant (Wilks' Lambda = 0.98, F= 93.86, df= (3, 17711), p < 0.0001), though the percent of variance explained was even lower (Pillai's trace = 0.016, p < 0.0001) than in the analysis including all levels (Pillai's trace = 0.022, p < 0.0001). Population density explains approximately 2% of the variance in academic achievement, academic environment, and at-risk behaviors. Gender explains 4% (Pillai's trace = 0.044, p < 0.0001) of the variance, and race explains 14% (Pillai's trace = 0.145, p < 0.0001) of the variance.



MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall V7202 Effect
H = Type III SSCP Matrix for V7202
E = Error SSCP Matrix

S=1 M=0.5 N=8854.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.95576646	273.23	3	17711	<.0001
Pillai's Trace	0.04423354	273.23	3	17711	<.0001
Hotelling-Lawley Trace	0.04628070	273.23	3	17711	<.0001
Roy's Greatest Root	0.04628070	273.23	3	17711	<.0001

Figure 5. MANOVA Output for the Null Hypothesis for Gender

The Type III Sum of Squares tables for each factor (Figure 16 - Figure 18, Appendix B) indicate that gender (p < 0.0001) and ethnicity (p < 0.0001) were statistically significantly related to academic achievement, though population density (p = 0.1506) was not. Ethnicity (p < 0.0001) and population density (p < 0.0001) were statistically significantly related to academic environment, which gender was not (p = 0.1150). Only population density (p < 0.0001) was shown to be statistically significantly related to Factor 3 (delinquency/at-risk behavior). Gender (p = 0.6338) and race (p = 0.5734) were not statistically significantly related to at-risk behaviors.

Post-hoc analysis by Tukey's Studentized Range Test was performed using a Bonferroni correction for multiple testing. Bonferroni correction was performed by dividing the desired level of statistical significance (α = 0.05) by the number of tests to be performed (21). The corrected alpha of 0.0024 was specified in a MEANS statement along with the TUKEY and CLDIFF (confidence levels of the estimated means) options. While the MEANS statement allows us to specify a different alpha level than the MANOVA analysis, it does not adjust for other terms in the model. The LSMEANS statement should be used to explore the adjusted means. The post-hoc analysis is summarized in Table 2.

Dep. Variable	Indep. Variable	Comparison	Difference Between Means	99.76% CI	Sig. *
Factor 1	Gender	Female – Male	0.420	0.375, 0.464	***
Factor 2	Gender	Female – Male	0.022	-0.020, 0.065	
Factor 3	Gender	Female – Male	0.006	-0.039, 0.052	
Factor 1	Race	White – Black	0.048	-0.027, 0.125	
		White – Hispanic	0.132	0.072, 0.192	***
		Black – Hispanic	0.084	-0.003, 0.170	
Factor 2	Race	White – Black	0.457	0.384, 0.530	***
		White – Hispanic	0.853	0.796, 0.910	***

Dep. Variable	Indep. Variable	Comparison	Difference Between Means	99.76% CI	Sig. *
		Black – Hispanic	0.396	0.313, 0.479	***
Factor 3	Race	Black – Hispanic	0.017	-0.017, 0.106	
		White – Black	-0.016	-0.093, 0.062	
		White – Hispanic	0.002	-0.059, 0.063	
Factor 1	Pop_density	High – Moderate	-0.014	-0.071, 0.043	
		High – Low	0.005	-0.063, 0.073	
		Moderate – Low	0.019	-0.044, 0.082	
Factor 2	Pop_density	High – Moderate	0.064	0.009, 0.118	***
		High – Low	0.158	0.091, 0.221	***
		Moderate – Low	0.092	0.032, 0.152	***
Factor 3	Pop_density	High – Moderate	0.026	-0.032, 0.084	
		High – Low	-0.064	-0.134, 0.006	
		Moderate – Low	-0.090	-0.154, -0.026	***

* Comparisons significant at the α = 0.0024 level are indicated by ***.

Table 2. Post-hoc Analysis Summary.

While statistically significant, the effect of population density was very small across Factors 2 and 3. While LSMEANS plots for population density and the academic environment factor shows a small, positive linear trend of increasing factor score with increasing population density, the LSMEANS plot for at-risk behaviors shows a U-shaped trend (Figure 6) that may be interesting to explore in future analyses. Please note that the difference between means for high-density and moderate-density was not shown to be statistically significant.

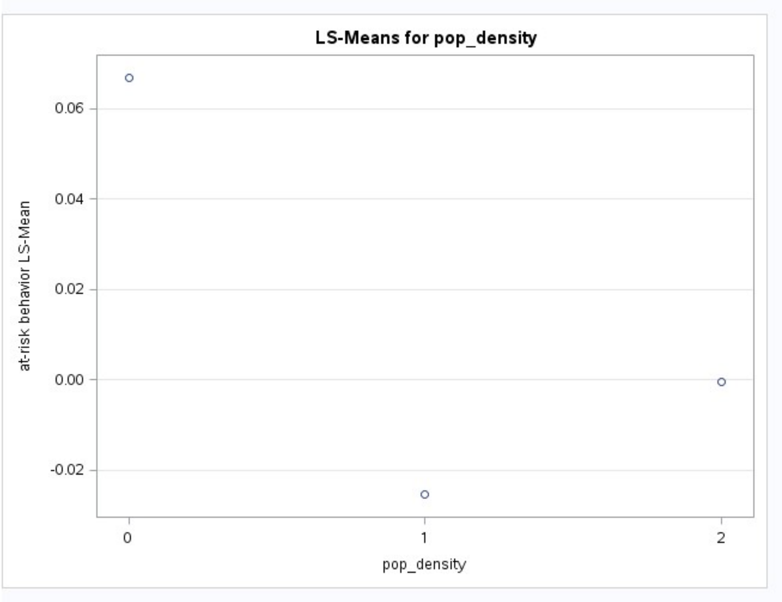


Figure 6. LSMeans Plot for At-Risk Behaviors (Factor 3) vs Population Density

CONCLUSION

This multivariate analysis of variance explored the relationships between three demographic independent variables and factors representing academic success, academic environment, and at-risk behaviors. This analysis has shown that the academic environment for 8th and 10th grade students, as described by Factor 2—which had high loadings for the variables for parental education, college preparatory program, and remedial schooling (negative loading)—varies significantly by race. The mean factor score for white students is highest, while Hispanic students have the lowest mean score for this factor. Population density explained a very low amount of the variance (2%) and should not be included in future exploratory or general analyses, as post-hoc analysis is strongly penalized for unnecessary groupings during multiple testing correction of the significance level.

REFERENCES

Carvalho, R. (2016). Gender differences in academic achievement: The mediating role of personality. *Personality and Individual Differences, 94*, 54-58. doi:10.1016/j.paid.2016.01.011

Johnston, L. D., Bachman, J. G., O'Malley, P. M., Schulenberg, J. E., & Miech, R. A. (2016). Codebook. *Monitoring the Future: A Continuing Study of American Youth (8th- and 10th-Grade Surveys), 2016.* Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Johnston, L. D., Bachman, J. G., O'Malley, P. M., Schulenberg, J. E., & Miech, R. A. (2017, October 26). *Monitoring the Future: A Continuing Study of American Youth (8th- and 10th-Grade Surveys), 2016*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor]. doi:https://doi.org/10.3886/ICPSR36799.v1

Johnston, L. D., O'Malley, P. M., Miech, R. A., Bachman, J. G., & Schulenberg, J. E. (2017, January). *Monitoring the Future national survey results on drug use, 1975-2016: Overview, key findings on adolescent drug use.* Ann Arbor, MI: Institute for Social Research, The University of Michigan .

Miller, P., & Vortruba-Drzal, E. (2015). Urbanicity moderates associations between family income and adolescent aademic ahievement. *Rural Sociology, 80*(3), 362-386. doi:10.1111/ruso.12067

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer’s MAP test. *Behavior Research Methods, Instruments & Computers, 32*(3), pp. 396-402.

Voyer, D., Voyer, S., & Hinshaw, S. (2014). Gender Differences in Scholastic Achievement: A Meta-Analysis. *Psychological Bulletin, 140*(4), 1174–1204. doi:10.1037/a0036620

ACKNOWLEDGMENTS

Appreciation is expressed to Dr. AnnMaria De Mars for comments on an earlier draft of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stefanie N. Kairs
10919 Parkdale Ave.
San Diego, CA 92126
stefanie.ness@gmail.com

APPENDIX A: SAS STUDIO CODE

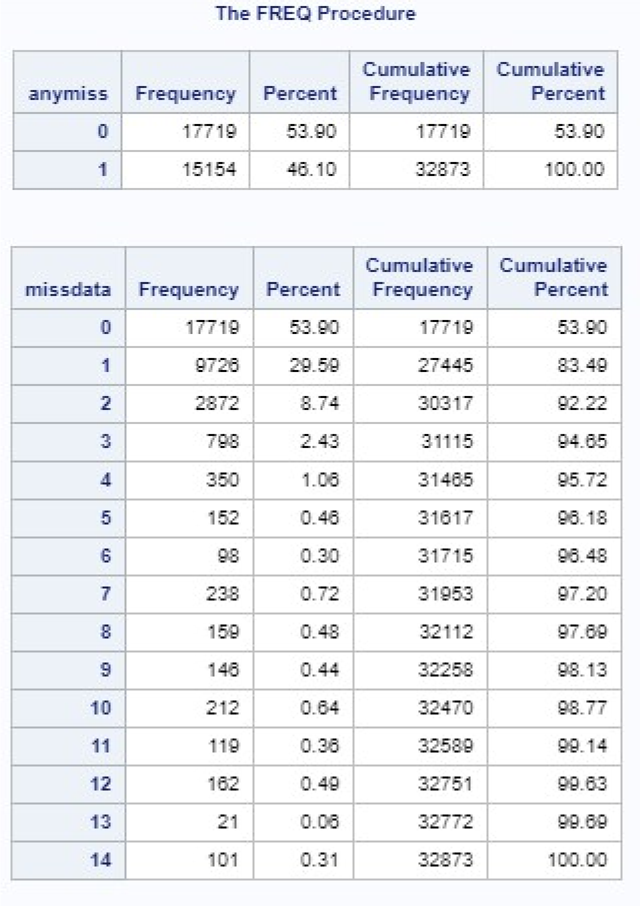
The SAS Studio code used to generate the analyses and figures for this paper is:

```
libname mydata "/courses/d82c65e5ba27fe300/c_7153/" access=readonly;
```

```
proc contents data=mydata.da36799p1;
run;

data fa_data;
  set mydata.da36799p1;
  keep V7102 V7107 V7215 V7216 V7221 V7222 V7223 V7226 V7231 V7232 V7233
       V7253 V7331 V7334 v501 v508 v509 v7202 v1070;
run;

/* recoding and parceling */
data fa_data_parcels;
  set fa_data;
  * recode for college prep = 1, no = 0;
  if v7222 > 1 then
     v7222=0;
  *recode to move "don't know" to missing (.);
  if v7215 > 6 then
     v7215=.;
  if v7216 > 6 then
     v7216=.;
  *parent education 7215 + 7216;
  parent_ed=v7215 + v7216;
  *remedial school;
  rem_school=v7232 + v7233;
  *population density <- for later MANOVA
    per codebook, additive combination of v508 and v509 yield
    0 = lowest density, 1 = moderate density, 2 = highest density;
  * using '+' to drop NA values, assuming NA = 0 is misleading;
  pop_density=v508 + v509;
  label parent_ed="Parental education" rem_school="Remedial schooling"
    pop_density="Population density";
  drop v508 v509 v7215 v7216 v7232 v7233;
run;

title " Descriptive statistics for numeric variables & pop_density";

proc means data=fa_data_parcels n nmiss std min mean median max;
  var V7102 V7107 V7221 V7222 V7223 V7226 V7231 V7253 V7331 V7334 parent_ed
    rem_school pop_density;
    *demographics removed except pop_density;
run;

title;
proc univariate data=fa_data_parcels;
  var V7102 V7107 V7221 V7222 V7223 V7226 V7231 V7253 V7331 V7334 parent_ed
    rem_school;
  histogram / normal kernel;
run;

proc freq data=fa_data_parcels;
  tables v501 v7202 v1070 pop_density;
run;

/* Missing data check */
Data check;
  set fa_data_parcels;
  array chckmiss{*} V7102 V7107 V7221 V7222 V7223 V7226 V7231 V7253 V7331
```

```
    V7334  parent_ed rem_school v7202 v1070 pop_density;
  missdata=0;

  do i=1 to dim(chckmiss);
    if chckmiss{i}=. then missdata=missdata + 1;
  end;

  if missdata > 0 then
     anymiss=1;
  else
     anymiss=0;
run;

/*Check for missing data differences by grade*/
proc freq data=check;
  tables anymiss missdata;
run;

proc freq data=check;
  tables anymiss*(v501 v7202 v1070) / chisq;
run;

proc ttest data=check;
  class v501;
  var missdata;
run;

/* Using settings from best FA from exploratory analysis */
Title " Factor Analysis: PCA varimax, parcels";
proc factor data=fa_data_parcels rotate=varimax reorder scree
    nfactors=3 out=fa_scored;
  var V7102 V7107 V7221 V7222 V7223 V7226 V7231 V7253 V7331 V7334 parent_ed
    rem_school;
run;

Data manova_data;
  set fa_scored;
  if factor1 ne .;
  label factor1="achievement" factor2="environment" factor3="at-risk
    behavior";
run;

/*Are the DVs correlated? */
proc corr data=manova_data;
  var factor1 - factor3;
run;

/* MANOVA factors1-3 by gender, race and pop density*/
proc glm data=manova_data plots=all;
  class v7202 v1070 pop_density;
  model factor1 - factor3=v7202 v1070 pop_density;
  CONTRAST 'High-density vs rest' pop_density 1 1 -2;
  Manova h=_all_ / printe printh;
  lsmeans v7202 v1070 pop_density;
  means v7202 v1070 pop_density/ alpha=0.0024 cldiff tukey;
run;
```

APPENDIX B: SUPPLEMENTAL FIGURES

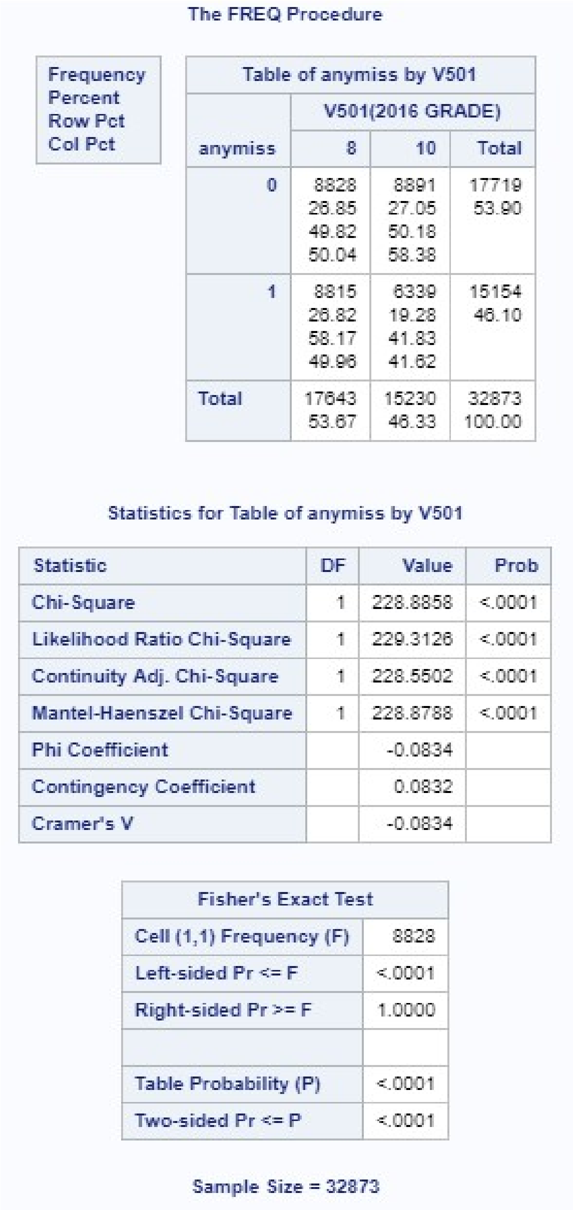


The FREQ Procedure

anymiss	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	17719	53.90	17719	53.90
1	15154	46.10	32873	100.00

missdata	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	17719	53.90	17719	53.90
1	9726	29.59	27445	83.49
2	2872	8.74	30317	92.22
3	798	2.43	31115	94.65
4	350	1.06	31465	95.72
5	152	0.46	31617	96.18
6	98	0.30	31715	96.48
7	238	0.72	31953	97.20
8	159	0.48	32112	97.69
9	146	0.44	32258	98.13
10	212	0.64	32470	98.77
11	119	0.36	32589	99.14
12	162	0.49	32751	99.63
13	21	0.06	32772	99.69
14	101	0.31	32873	100.00

Figure 7. Missing Data Evaluation



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of anymiss by V501			
	V501(2016 GRADE)		
anymiss	8	10	Total
0	8828 26.85 49.82 50.04	8891 27.05 50.18 58.38	17719 53.90
1	8815 26.82 58.17 49.96	6339 19.28 41.83 41.62	15154 46.10
Total	17643 53.67	15230 46.33	32873 100.00

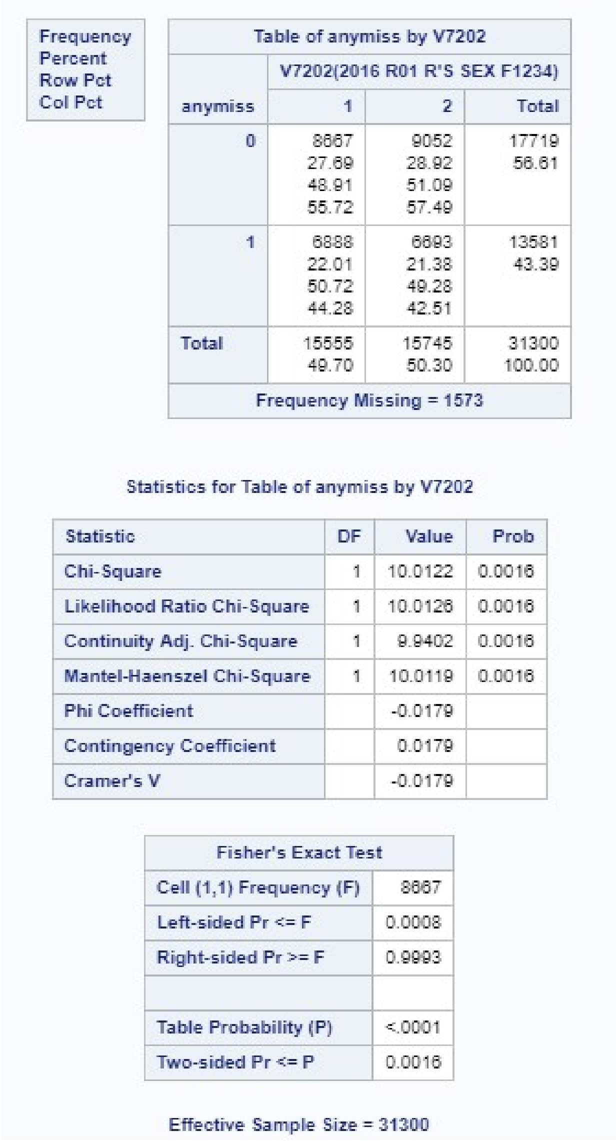
Statistics for Table of anymiss by V501

Statistic	DF	Value	Prob
Chi-Square	1	228.8858	<.0001
Likelihood Ratio Chi-Square	1	229.3126	<.0001
Continuity Adj. Chi-Square	1	228.5502	<.0001
Mantel-Haenszel Chi-Square	1	228.8788	<.0001
Phi Coefficient		-0.0834	
Contingency Coefficient		0.0832	
Cramer's V		-0.0834	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	8828
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 32873

Figure 8. Missing Values by Grade Level



Frequency Percent Row Pct Col Pct

Table of anymiss by V7202			
	V7202(2016 R01 R'S SEX F1234)		
anymiss	1	2	Total
0	8667 27.69 48.91 55.72	9052 28.92 51.09 57.49	17719 56.61
1	6888 22.01 50.72 44.28	6693 21.38 49.28 42.51	13581 43.39
Total	15555 49.70	15745 50.30	31300 100.00
Frequency Missing = 1573			

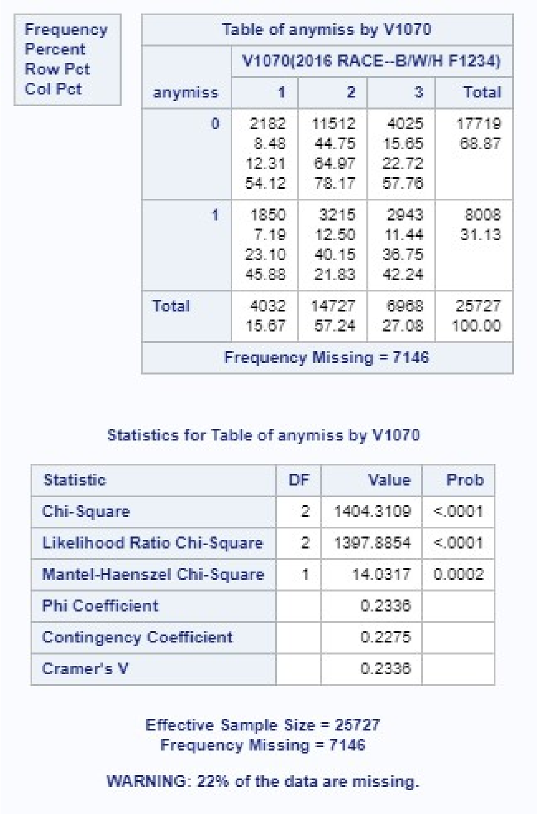
Statistics for Table of anymiss by V7202

Statistic	DF	Value	Prob
Chi-Square	1	10.0122	0.0016
Likelihood Ratio Chi-Square	1	10.0126	0.0016
Continuity Adj. Chi-Square	1	9.9402	0.0016
Mantel-Haenszel Chi-Square	1	10.0119	0.0016
Phi Coefficient		-0.0179	
Contingency Coefficient		0.0179	
Cramer's V		-0.0179	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	8667
Left-sided Pr <= F	0.0008
Right-sided Pr >= F	0.9993
Table Probability (P)	<.0001
Two-sided Pr <= P	0.0016

Effective Sample Size = 31300

Figure 9. Missing Values by Gender



Frequency Percent Row Pct Col Pct

Table of anymiss by V1070

	V1070(2016 RACE--B/W/H F1234)			
anymiss	1	2	3	Total
0	2182 8.48 12.31 54.12	11512 44.75 64.97 78.17	4025 15.65 22.72 57.76	17719 68.87
1	1850 7.19 23.10 45.88	3215 12.50 40.15 21.83	2943 11.44 36.75 42.24	8008 31.13
Total	4032 15.67	14727 57.24	6968 27.08	25727 100.00

Frequency Missing = 7146

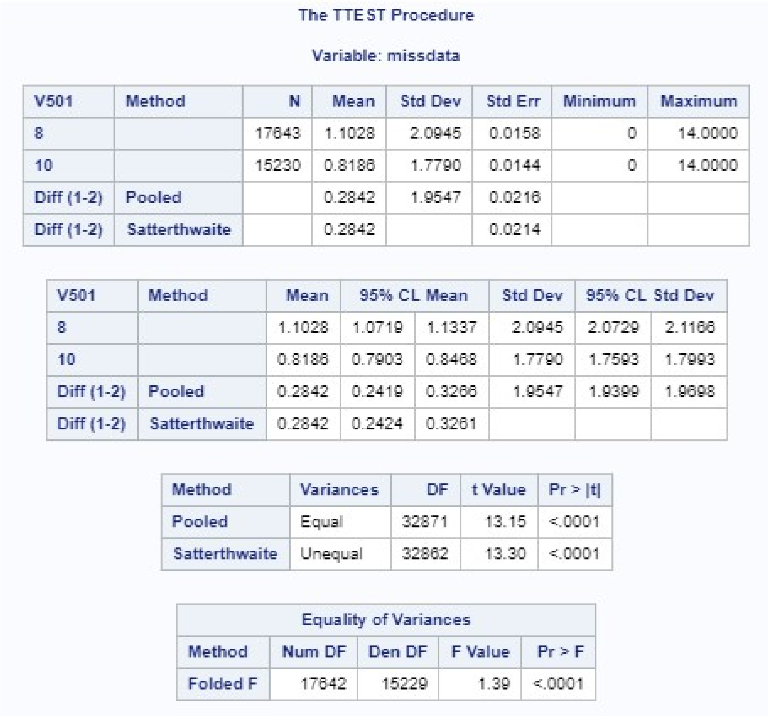
Statistics for Table of anymiss by V1070

Statistic	DF	Value	Prob
Chi-Square	2	1404.3109	<.0001
Likelihood Ratio Chi-Square	2	1397.8854	<.0001
Mantel-Haenszel Chi-Square	1	14.0317	0.0002
Phi Coefficient		0.2336	
Contingency Coefficient		0.2275	
Cramer's V		0.2336	

Effective Sample Size = 25727
Frequency Missing = 7146

WARNING: 22% of the data are missing.

Figure 10. Missing Values by Race



The TTEST Procedure

Variable: missdata

V501	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
8		17643	1.1028	2.0945	0.0158	0	14.0000
10		15230	0.8186	1.7790	0.0144	0	14.0000
Diff (1-2)	Pooled		0.2842	1.9547	0.0216		
Diff (1-2)	Satterthwaite		0.2842		0.0214		

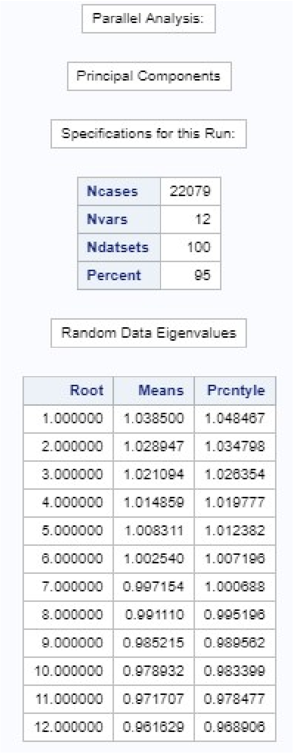
V501	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
8		1.1028	1.0719	1.1337	2.0945	2.0729	2.1166
10		0.8186	0.7903	0.8468	1.7790	1.7593	1.7993
Diff (1-2)	Pooled	0.2842	0.2419	0.3266	1.9547	1.9399	1.9698
Diff (1-2)	Satterthwaite	0.2842	0.2424	0.3261			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	32871	13.15	<.0001
Satterthwaite	Unequal	32862	13.30	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	17642	15229	1.39	<.0001

Figure 11. T-test for Number of Missing Values by Grade Level



Parallel Analysis:

Principal Components

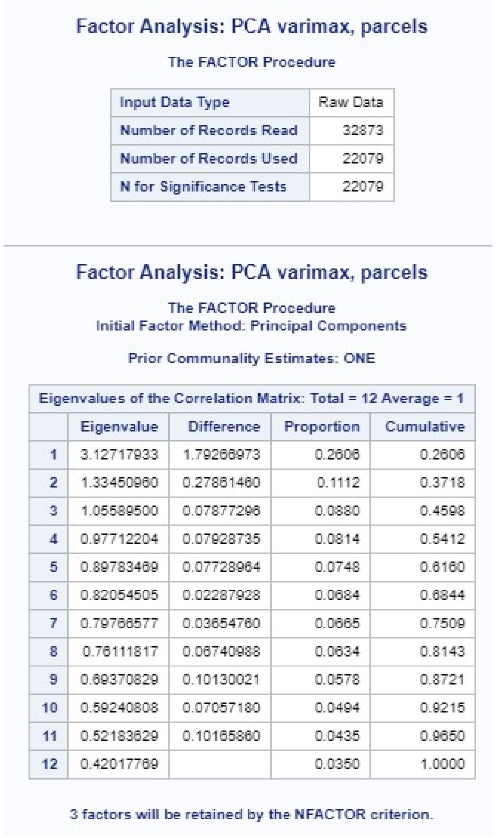
Specifications for this Run:

Ncases	22079
Nvars	12
Ndatsets	100
Percent	95

Random Data Eigenvalues

Root	Means	Prcntyle
1.000000	1.038500	1.048467
2.000000	1.028947	1.034798
3.000000	1.021094	1.026354
4.000000	1.014859	1.019777
5.000000	1.008311	1.012382
6.000000	1.002540	1.007196
7.000000	0.997154	1.000688
8.000000	0.991110	0.995196
9.000000	0.985215	0.989562
10.000000	0.978932	0.983399
11.000000	0.971707	0.978477
12.000000	0.961629	0.968906

Figure 12. Parallel Analysis Criterion Script Output



Factor Analysis: PCA varimax, parcels

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	32873
Number of Records Used	22079
N for Significance Tests	22079

Factor Analysis: PCA varimax, parcels

The FACTOR Procedure
Initial Factor Method: Principal Components

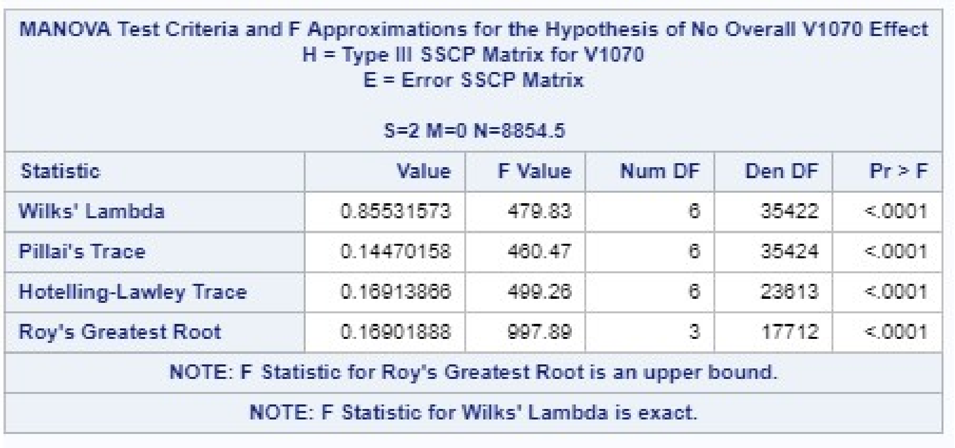
Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 12 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.12717933	1.79266973	0.2606	0.2606
2	1.33450960	0.27861460	0.1112	0.3718
3	1.05589500	0.07877296	0.0880	0.4598
4	0.97712204	0.07928735	0.0814	0.5412
5	0.89783469	0.07728964	0.0748	0.6160
6	0.82054505	0.02287928	0.0684	0.6844
7	0.79766577	0.03654760	0.0665	0.7509
8	0.76111817	0.06740988	0.0634	0.8143
9	0.69370829	0.10130021	0.0578	0.8721
10	0.59240808	0.07057180	0.0494	0.9215
11	0.52183629	0.10165860	0.0435	0.9650
12	0.42017769		0.0350	1.0000

3 factors will be retained by the NFACTOR criterion.

Figure 13. Factor Analysis: Eigenvalues of Correlation Matrix

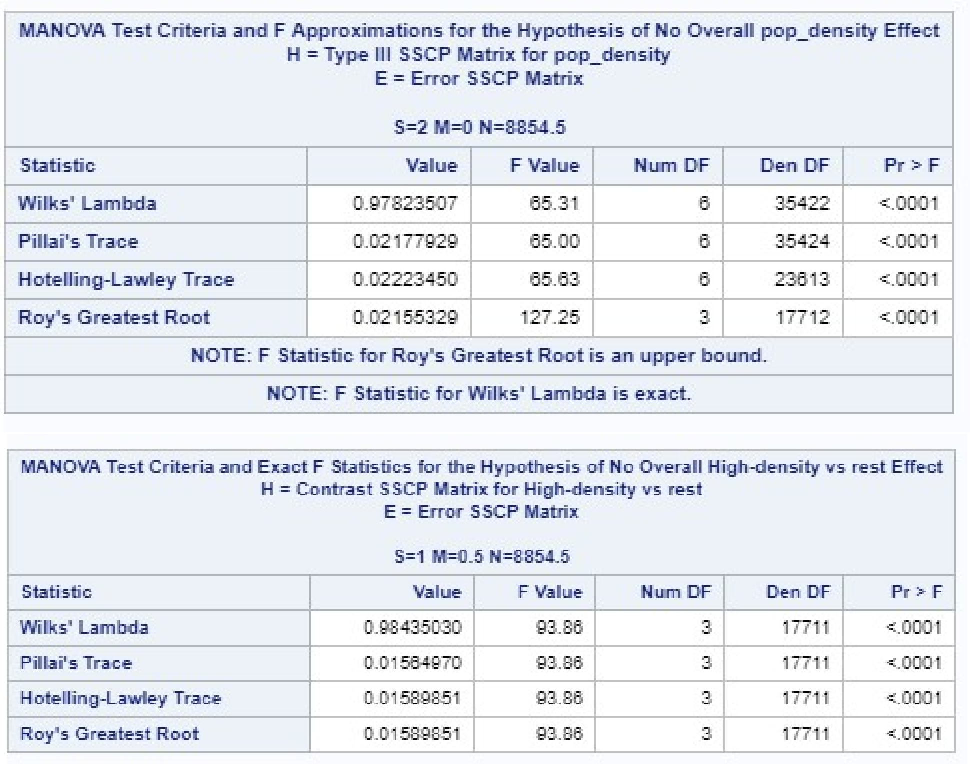


MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall V1070 Effect
H = Type III SSCP Matrix for V1070
E = Error SSCP Matrix

S=2 M=0 N=8854.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.85531573	479.83	6	35422	<.0001
Pillai's Trace	0.14470158	460.47	6	35424	<.0001
Hotelling-Lawley Trace	0.16913866	499.26	6	23613	<.0001
Roy's Greatest Root	0.16901888	997.89	3	17712	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Figure 14. MANOVA Output for the Null Hypothesis for Race



MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall pop_density Effect
H = Type III SSCP Matrix for pop_density
E = Error SSCP Matrix

S=2 M=0 N=8854.5

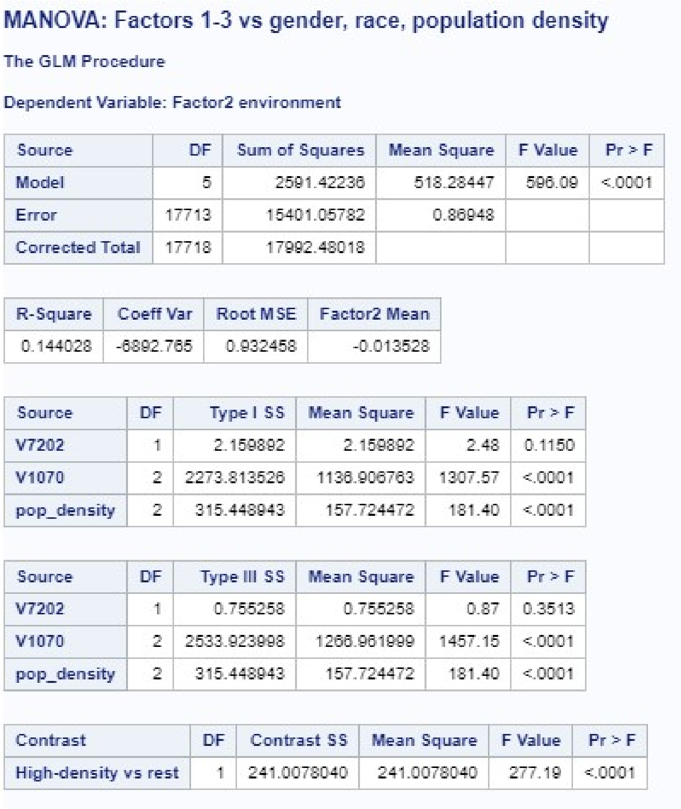
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.97823507	65.31	6	35422	<.0001
Pillai's Trace	0.02177929	65.00	6	35424	<.0001
Hotelling-Lawley Trace	0.02223450	65.63	6	23613	<.0001
Roy's Greatest Root	0.02155329	127.25	3	17712	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall High-density vs rest Effect
H = Contrast SSCP Matrix for High-density vs rest
E = Error SSCP Matrix

S=1 M=0.5 N=8854.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.98435030	93.86	3	17711	<.0001
Pillai's Trace	0.01564970	93.86	3	17711	<.0001
Hotelling-Lawley Trace	0.01589851	93.86	3	17711	<.0001
Roy's Greatest Root	0.01589851	93.86	3	17711	<.0001

Figure 15. MANOVA Output for the Null Hypotheses for Population Density



MANOVA: Factors 1-3 vs gender, race, population density

The GLM Procedure

Dependent Variable: Factor2 environment

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2591.42236	518.28447	596.09	<.0001
Error	17713	15401.05782	0.86948		
Corrected Total	17718	17992.48018			

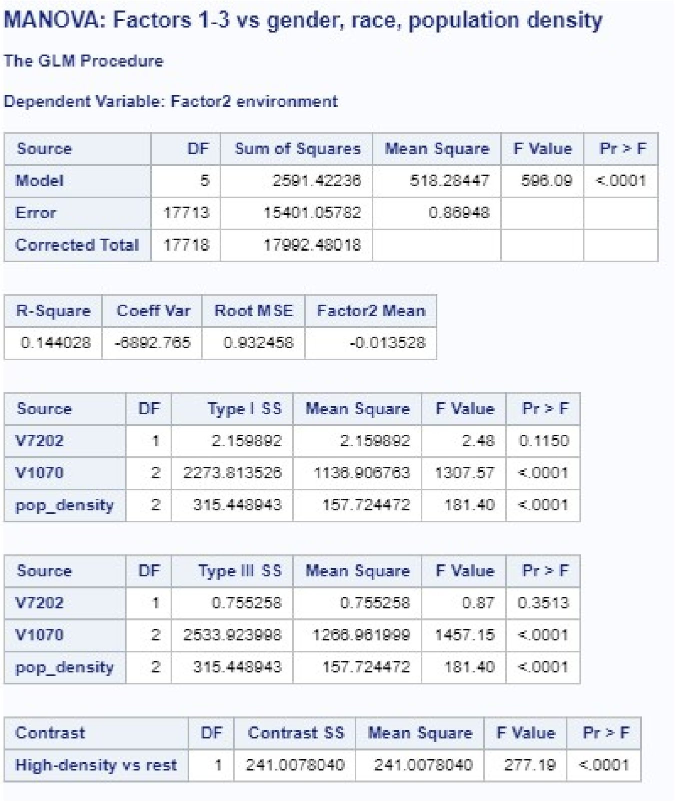
R-Square	Coeff Var	Root MSE	Factor2 Mean
0.144028	-6892.765	0.932458	-0.013528

Source	DF	Type I SS	Mean Square	F Value	Pr > F
V7202	1	2.159892	2.159892	2.48	0.1150
V1070	2	2273.813526	1136.906763	1307.57	<.0001
pop_density	2	315.448943	157.724472	181.40	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
V7202	1	0.755258	0.755258	0.87	0.3513
V1070	2	2533.923998	1266.961999	1457.15	<.0001
pop_density	2	315.448943	157.724472	181.40	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
High-density vs rest	1	241.0078040	241.0078040	277.19	<.0001

Figure 16. MANOVA Output Factor 1 (Achievement)



MANOVA: Factors 1-3 vs gender, race, population density

The GLM Procedure

Dependent Variable: Factor2 environment

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2591.42236	518.28447	596.09	<.0001
Error	17713	15401.05782	0.86948		
Corrected Total	17718	17992.48018			

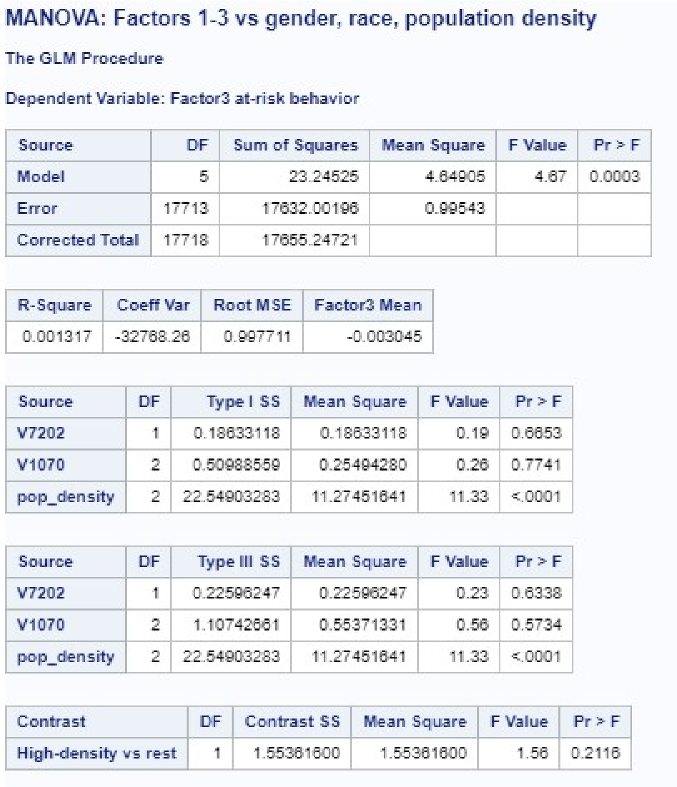
R-Square	Coeff Var	Root MSE	Factor2 Mean
0.144028	-6892.765	0.932458	-0.013528

Source	DF	Type I SS	Mean Square	F Value	Pr > F
V7202	1	2.159892	2.159892	2.48	0.1150
V1070	2	2273.813526	1136.906763	1307.57	<.0001
pop_density	2	315.448943	157.724472	181.40	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
V7202	1	0.755258	0.755258	0.87	0.3513
V1070	2	2533.923998	1266.961999	1457.15	<.0001
pop_density	2	315.448943	157.724472	181.40	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
High-density vs rest	1	241.0078040	241.0078040	277.19	<.0001

Figure 17. MANOVA Output Factor 2 (Environment)



MANOVA: Factors 1-3 vs gender, race, population density

The GLM Procedure

Dependent Variable: Factor3 at-risk behavior

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	23.24525	4.64905	4.67	0.0003
Error	17713	17632.00196	0.99543		
Corrected Total	17718	17655.24721			

R-Square	Coeff Var	Root MSE	Factor3 Mean
0.001317	-32768.26	0.997711	-0.003045

Source	DF	Type I SS	Mean Square	F Value	Pr > F
V7202	1	0.18633118	0.18633118	0.19	0.6653
V1070	2	0.50988559	0.25494280	0.26	0.7741
pop_density	2	22.54903283	11.27451641	11.33	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
V7202	1	0.22596247	0.22596247	0.23	0.6338
V1070	2	1.10742661	0.55371331	0.56	0.5734
pop_density	2	22.54903283	11.27451641	11.33	<.0001

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
High-density vs rest	1	1.55361600	1.55361600	1.56	0.2116

Figure 18. MANOVA Output Factor 3 (At-Risk Behavior)