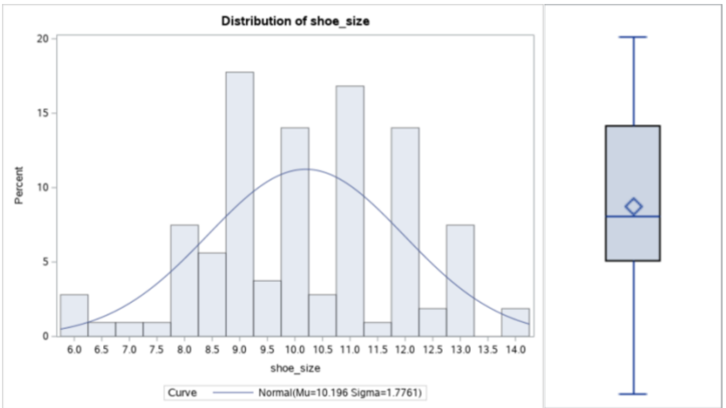
Introduction

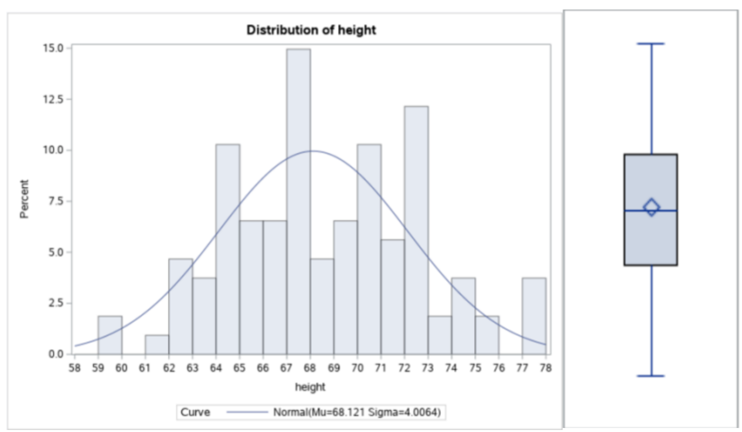
My project will center around determining whether factors such as race, shoe size, height and gender contribute to a person's preferred sport, as well as finding out if there is a relationship between a person's height and his/her shoe size. The quantitative variables, shoe size and height, will be analyzed in order to see if there is any relationship between heights and shoe sizes. The qualitative variables (gender, race and place of origin) will shine some light on if there is any relationship between a certain demographic and any sport. The linear regression section will focus on the quantitative variables and gender to model the relationship between shoe size, gender and height. With the variables being analyzed, it is clear to see that there are multiple conclusions that can be drawn from this dataset.

Descriptive Statistics

Quantitative Data



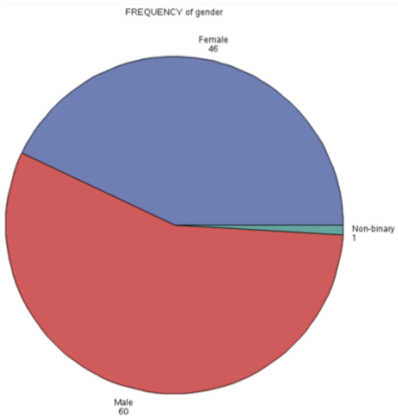
For the height, the data seems to be multimodal and somewhat symmetric (skewness = 0.12537703). According to the skewness and the simple statistics which tells us that the mean > median, the data seems to have a slight right skew. The box plot tells us there are no outliers and the data seems to be generally symmetric, so the best measures of center and spread would be mean, variance and standard deviation since they take into account all of the data points. For the shoe sizes, the data seems to be similar to the height in the sense that it is also multimodal and somewhat symmetric (skewness = -0.1062754). Also like the height data, there seems to be a slight right skew according to the general shape of the histogram and mean being greater than median. There are also no outliers in this data and there seems to be symmetry so we will use the same measures of center and spread: mean, variance and standard deviation.

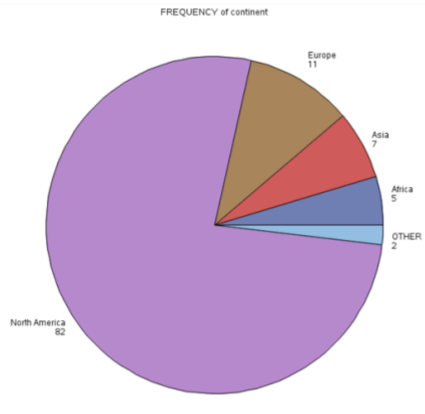


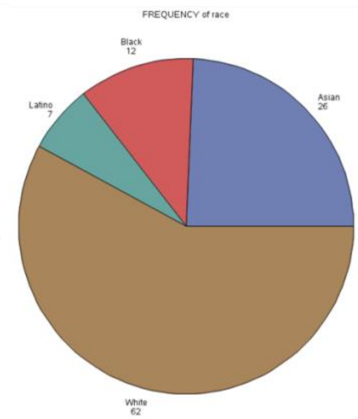
Variable	Mean	Median	Std Dev	Quartile Range	Maximum	Minimum
shoe_size	10.20	10.00	1.78	3.00	14.00	6.00
height	68.12	68.00	4.01	6.00	77.00	59.00

Qualitative Data

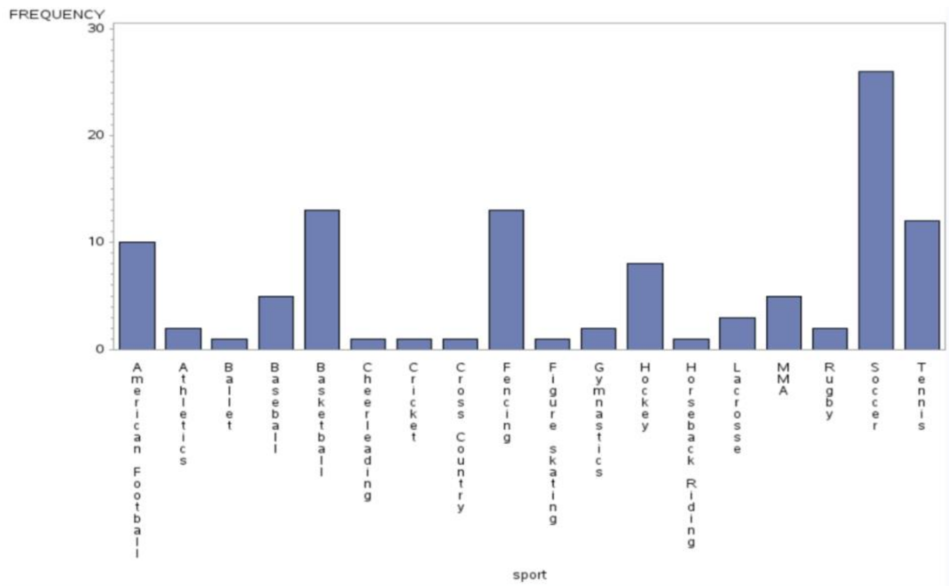
After collecting the data and analyzing the categorical variables, it is clear to see that certain groups dominate. For example, in the continent and race variables, North America and White account for about 76.7% and 57.9% of the data. The gender variable seems to be more evenly spread out for the most part (Male has the most with 56.1%).







For the sport categorical variable, there are multiple sports that were only picked once out 107 times. This will probably affect the reliability of the data when it comes to analyzing and



determining if there is a relationship between two variables. Tests involving the sports variable such as the chi-square test for independence will need to be taken with a huge grain of salt. With that being said, I still ran the chi-square test just to see the values that would be yielded.

Statistics for Table of gender by sport

Statistic	DF	Value	Prob
Chi-Square	34	25.4730	0.8539
Likelihood Ratio Chi-Square	34	26.4932	0.8173
Mantel-Haenszel Chi-Square	1	2.8806	0.0897
Phi Coefficient		0.4879	
Contingency Coefficient		0.4385	
Cramer's V		0.3450	
WARNING: 83% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Statistics for Table of continent by sport

Statistic	DF	Value	Prob
Chi-Square	85	126.1178	0.0025
Likelihood Ratio Chi-Square	85	67.3717	0.9202
Mantel-Haenszel Chi-Square	1	0.1743	0.6763
Phi Coefficient		1.0857	
Contingency Coefficient		0.7355	
Cramer's V		0.4855	
WARNING: 94% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Statistics for Table of race by sport

Statistic	DF	Value	Prob
Chi-Square	51	60.4673	0.1710
Likelihood Ratio Chi-Square	51	65.8120	0.0794
Mantel-Haenszel Chi-Square	1	2.7290	0.0985
Phi Coefficient		0.7517	
Contingency Coefficient		0.6009	
Cramer's V		0.4340	
WARNING: 92% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

According to the chi square values, a person's favorite sport shares the strongest relationship with the continent they are from, followed by race then gender (126 > 60 > 25). As the warnings reiterate, chi square may not be a valid test due to the spread of the data. Since the continent categorical variable had the highest chi-squared, we used a two-way frequency tables to further analyze the relationship between continents and sports. North Americans seem to have dominant numbers in many sports such as fencing, baseball and soccer so these may be some popular sports in that region (76.7% of the surveys were filled by North Americans so again, the reliability of this data must be called into question). 80% of the Africans picked soccer, 43.8% of Asians picked fencing and 36.36% of Europeans picked basketball and soccer, so these sports may also be huge in those continents (Note: We also had significantly less observations from these regions as compared to North America).

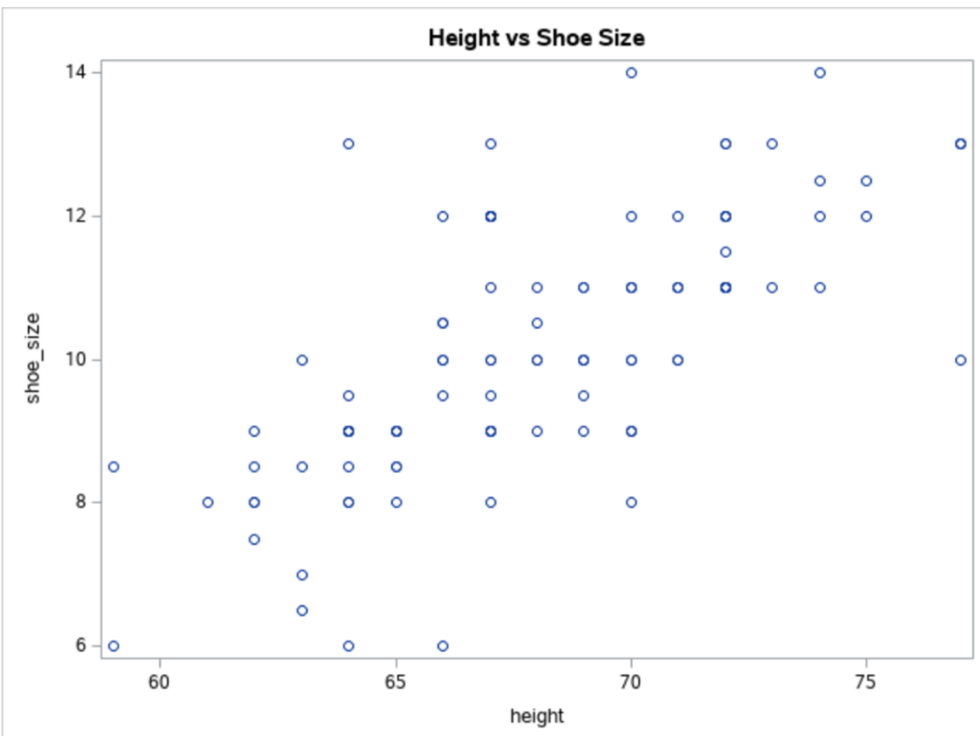
Table of continent by sport

	sport																		
continent	American Football	Athletics	Ballet	Baseball	Basketball	Cheerleading	Cricket	Cross Country	Fencing	Figure skating	Gymnastics	Hockey	Horseback Riding	Lacrosse	MMA	Rugby	Soccer	Tennis	Total
Africa	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 20.00 7.69	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 3.74 80.00 15.38	0 0.00 0.00 0.00	5 4.67
Asia	1 0.93 14.29 10.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 14.29 100.00	0 0.00 0.00 0.00	3 2.80 42.86 23.08	1 0.93 14.29 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 14.29 3.85	0 0.00 0.00 0.00	7 6.54
Australia	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 100.00 50.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93
Europe	0 0.00 0.00 0.00	1 0.93 9.09 50.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 3.74 36.36 30.77	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 9.09 50.00	4 3.74 36.36 15.38	1 0.93 9.09 8.33	11 10.28
North America	9 8.41 10.98 90.00	1 0.93 1.22 50.00	1 0.93 1.22 100.00	5 4.67 6.10 100.00	8 7.48 9.76 61.54	1 0.93 1.22 100.00	0 0.00 0.00 0.00	1 0.93 1.22 100.00	10 9.35 12.20 76.92	0 0.00 0.00 0.00	2 1.87 2.44 100.00	8 7.48 9.76 100.00	1 0.93 1.22 100.00	3 2.80 3.66 100.00	5 4.67 6.10 100.00	0 0.00 0.00 0.00	16 14.95 19.51 61.54	11 10.28 13.41 91.67	82 76.64
South America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.93 100.00 3.85	0 0.00 0.00 0.00	1 0.93
Total	10 9.35	2 1.87	1 0.93	5 4.67	13 12.15	1 0.93	1 0.93	1 0.93	13 12.15	1 0.93	2 1.87	8 7.48	1 0.93	3 2.80	5 4.67	2 1.87	26 24.30	12 11.21	107 100.00

<u>**Correlation**</u>

Pearson Correlation Coefficients, N = 107 Prob > |r| under H0: Rho=0		
	shoe_size	height
shoe_size	1.00000	0.69266 <.0001
height	0.69266 <.0001	1.00000

Since we only have two quantitative variables, there is only one correlation coefficient to find. After running the "PROC CORR" procedure on SAS to get the Pearson correlation, the correlation coefficient for our quantitative variables we get is 0.69266. This r value shows that there is a positive and moderately strong relationship between height and shoe sizes. So, as the height increases, shoe size is usually expected to increase as well. The scatter plot exhibits this somewhat strong relationship well:



Linear Regression Model

For the first linear regression, we analyzed only the quantitative variables, height and shoe size. We have a r^2 value of 0.4748. This coefficient of determination tells us that 47.48% of the variation in shoe size can be explained by the variation in height which is not particularly good when you're looking to build a strong linear regression model. The Analysis of Variance gives us a p-value of <.0001, letting us know that the null hypothesis is rejected, and the independent variable (height) is significant. The root MSE here is 1.28712. This is only useful when comparing to other models, so we will come back to this later on. The p-values for the parameter estimates are all below .0001 which is a good sign for the regression model since we're pretty sure that these values don't equal 0.

The REG Procedure
Model: MODEL1
Dependent Variable: shoe_size

Number of Observations Read	107
Number of Observations Used	107

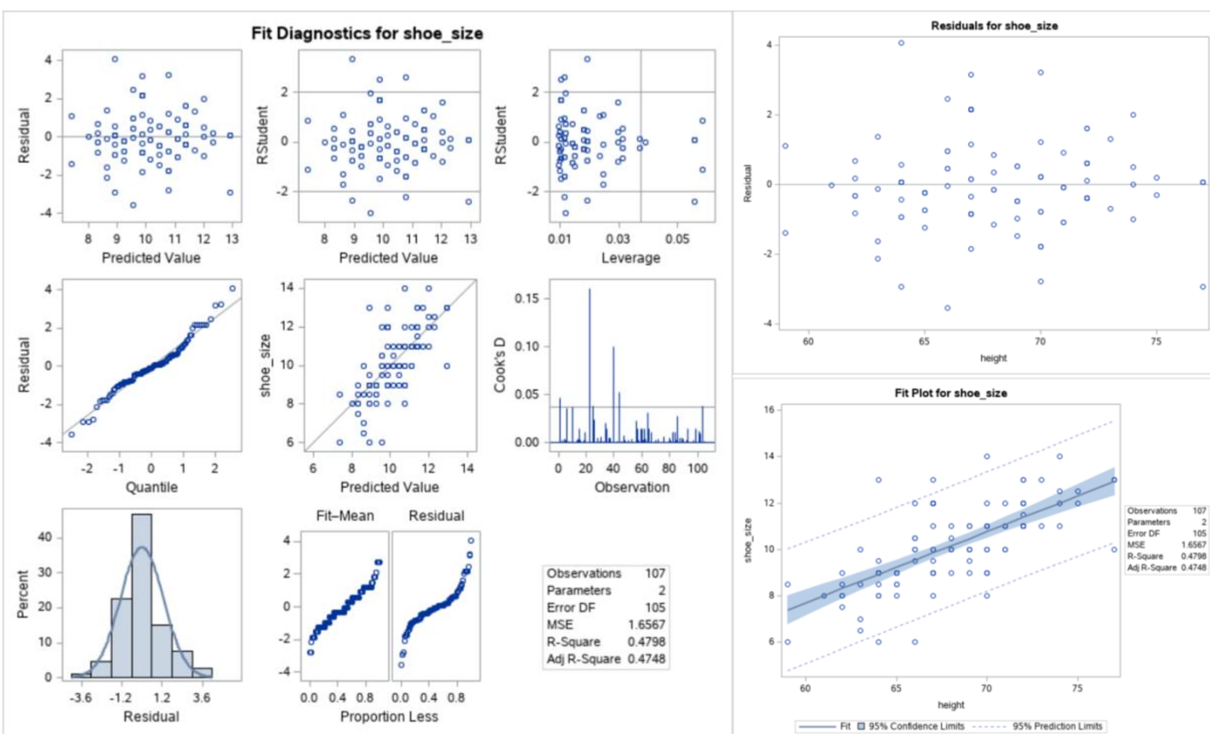
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	160.42626	160.42626	96.84	<.0001
Error	105	173.95224	1.65669		
Corrected Total	106	334.37850			

Root MSE	1.28712	R-Square	0.4798
Dependent Mean	10.19626	Adj R-Sq	0.4748
Coeff Var	12.62349		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-10.72154	2.12932	-5.04	<.0001
height	1	0.30707	0.03120	9.84	<.0001



From the parameter estimates, we get a regression equation of:

$$shoe\ size = 0.30707 * height - 10.72154.$$

The residual plot looks very good: It is not systematically high or low, it is centered on zero and no patterns seem to be forming so there is a randomness to the points. Based off of the Q-Q plot, the data seems to be normal, but the Cook's D plot reveals the presence of a few outliers. After taking everything into account, this regression equation seems to be average.

The REG Procedure
Model: MODEL1
Dependent Variable: shoe_size

Number of Observations Read	107
Number of Observations Used	107

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	169.47637	56.49212	35.29	<.0001
Error	103	164.90214	1.60099		
Corrected Total	106	334.37850			

Root MSE	1.26530	R-Square	0.5068
Dependent Mean	10.19626	Adj R-Sq	0.4925
Coeff Var	12.40948		

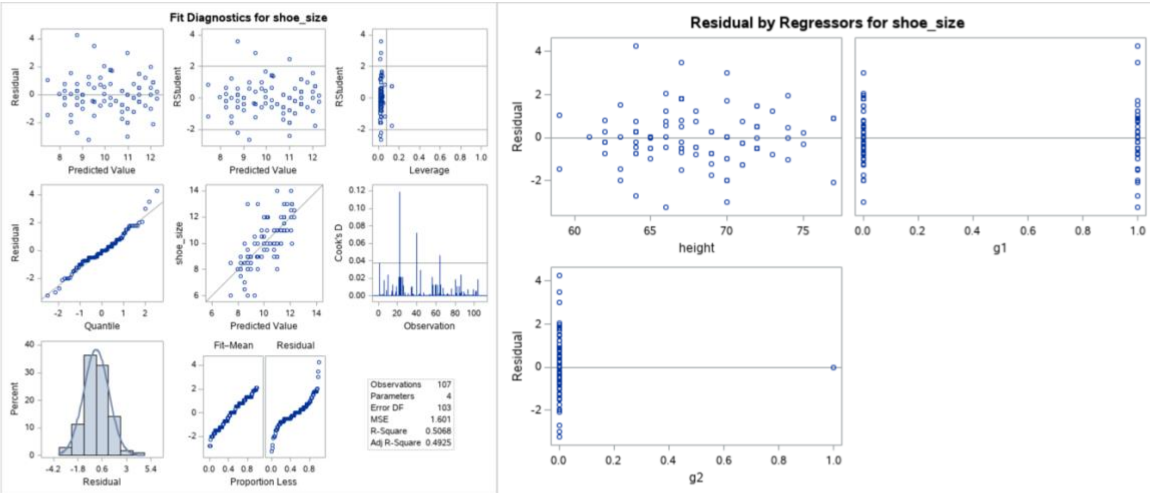
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.13159	2.58149	-2.76	0.0068
height	1	0.25890	0.03677	7.04	<.0001
g1	1	-0.70240	0.29637	-2.37	0.0196
g2	1	-0.69660	1.28933	-0.54	0.5902

The second regression utilized the categorical variable, gender, in addition to the quantitative variables. Since there were three genders selected, we had two dummy variables (g1 and g2) with 'Male' serving as the reference category. The r^2 value of 0.5068 so 50.68% of the variation in shoe size can be explained by the variation height, g1 and g2. While this is marginally better value than the one from the first regression, it is still not that good. The ANOVA p-value is also <.0001 so the null hypothesis is again rejected, and the independent variables are significant. The root MSE here is 1.26530 which is just marginally better than the 1.28712 from the first regression model. The p-values for the height (<.0001) and g1 (.0196) are below .05 but g2 is 0.5902. The g2 p-value is clearly not good but we will keep analyzing to see if the model is useful and we want to accept a lesser confidence level. From the parameter estimates, we also get the regression equation:

$$shoe\ size = 0.2589 * height - 0.7024 * g1 - 0.6966 * g2 - 7.13159$$

The residual plot is just as good as the one from the first model for the same reasons: It is not systematically high or low, it is centered on zero and there are no patterns forming so there is no predictable nature to the points. The Quantile-Quantile plot shows that there is definitely some normality and the Cook's D shows that three outliers exist.



The REG Procedure
Model: MODEL1
Dependent Variable: shoe_size

Number of Observations Read	107
Number of Observations Used	107

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	169.00904	84.50452	53.14	<.0001
Error	104	165.36947	1.59009		
Corrected Total	106	334.37850			

Root MSE	1.26099	R-Square	0.5054
Dependent Mean	10.19626	Adj R-Sq	0.4959
Coeff Var	12.36716		

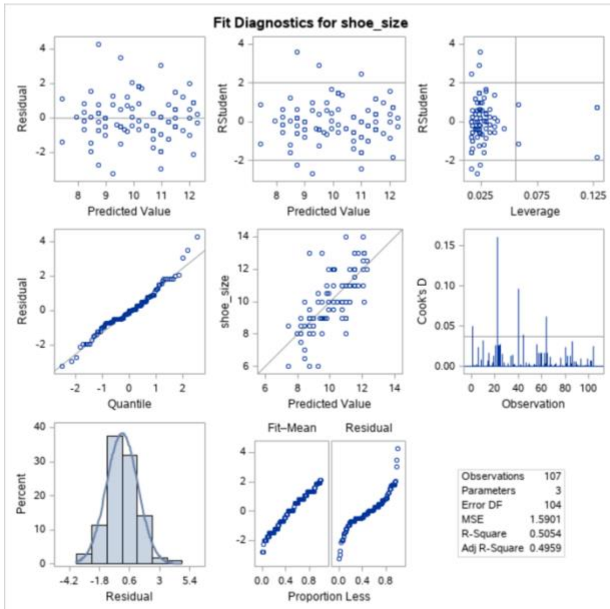
Parameter Estimates

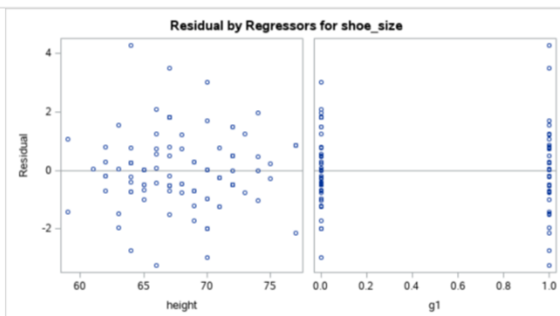
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.34389	2.54271	-2.89	0.0047
height	1	0.26177	0.03626	7.22	<.0001
g1	1	-0.67855	0.29206	-2.32	0.0221

The last regression model we will be looking involves the dependent variable shoe size, and the independent variables, height and only g1. 0.5054 was the r^2 value so 50.54% of the variation in shoe size can be explained by the variation height and g1 (marginally less than the second model but more than the first). The p-value for ANOVA is still <.0001 so everything is fine so far since the independent values are all relevant (none equal 0, null hypothesis rejected). The root MSE here is the best we've seen so far 1.26099. The p-values for height (<.0001) and g1 (.0221) are below .05 which is a good sign for the regression model and is a clear improvement on the previous one. In the same area, we are able to derive a regression equation:

$$shoe\ size = 0.26177 * height - 0.67855 * g1 - 7.34389$$

Much like its predecessors, the regression plot here is very impressive as the points are very well spread out with no patterns in sight, thus allowing there to be an element of randomness and unpredictability. The Quantile-Quantile plot supports the claim that the data is normally distributed, and the Cook's D plot tells us that there are four outliers.





These three regressions models were selected to be analyzed due to the fact that they possessed the highest adjusted r-squared value when we ran the "PROC REG" procedure with the ADJRSQ selection.

The REG Procedure
Model: MODEL1
Dependent Variable: shoe_size

Adjusted R-Square Selection Method

Number of Observations Read	107
Number of Observations Used	107

Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.4959	0.5054	height g1
3	0.4925	0.5068	height g1 g2
1	0.4748	0.4798	height
2	0.4699	0.4799	height g2
2	0.2554	0.2695	g1 g2
1	0.2505	0.2576	g1
1	-.0052	0.0043	g2

All regression models are very similar in that their values are very comparable, but out of the three, the third regression model (shoe size, height, g1) is the best for the following reasons: It has the second highest r^2 value (the higher adjusted r^2 and very close to being the highest r^2), the best root MSE, all its independent variables have good p-values, and the residual plot looks great. The other two models possess a lot of these traits, but this model just edges it when taking everything into account.

Supporting Documentation

SAS Code

All relevant SAS output already appears in the document.

```
/* DATA Step  */
data favoritesport;
infile "/home/u48688562/classwork/STAT 430 Project on Sports.csv" delimiter="," firstobs=3;
format continent race sport gender $30.;
input gender $ continent $ shoe_size race $ height sport $;
run;

/* Generating the Simple Statistics for Quantitative Variables*/
proc means data=favoritesport mean median stddev qrange max min maxdec=2;
var shoe_size height;
run;

/* Histograms for Quantitative Variables */
proc univariate data=favoritesport plot;
histogram shoe_size/ normal midpoints=(6 to 14 by .5);
histogram height/ normal endpoints=(58 to 78 by 1);
run;

/* Frequency Tables for Qualitative Variables */
proc freq data=favoritesport;
tables gender continent race sport;
run;

/* Two-Way Frequency Tables for Qualitative Variables */
/* Chi Square Test to see if there is a relationship between variables */
proc freq data=favoritesport;
tables gender*sport continent*sport race*sport/chisq;
run;

/* Pie Chart for the frequencies of Qualitative Variables */
proc gchart data=favoritesport;
pie gender;
pie continent;
pie race;
vbar sport;
run;


/* Check for Correlation */
proc corr data=favoritesport;
var shoe_size height;
run;

/* Scatter Plot */
proc sgplot data=favoritesport;
title "Height vs Shoe Size";
scatter y=shoe_size x=height;
run;

/* Linear Regression model with just Quantitative Variables */
proc reg data=favoritesport;
model shoe_size = height;
run;

/* DATA Step to include dummy variables for gender */
data favoritesport1;
set favoritesport;
if gender = 'Male' then do;
g1 = 0; g2 = 0;
end;
else if gender = 'Female' then do;
g1 = 1; g2 = 0;
end;
else do;
g1 = 0; g2 = 1;
end;
run;

/* Linear Regression model to pick models wit adjusted r squared*/
proc reg data=favoritesport1;
model shoe_size = height g1 g2/selection=adjrsq;
run;

/* Linear Regression model with Quantitative Variables and dummy variables for gender*/
proc reg data=favoritesport1;
model shoe_size = height g1 g2;
run;
```

Survey

https://docs.google.com/forms/d/e/1FAIpQLSe3t_hMSQC4HMSc4VQvKRh0J1A9MrgZGOMynJafK0oyJVSiNQ/viewform

STAT 430 Project

* Required

What is your gender? *

- Female
- Male
- Other:

Which continent are you from? *

- North America
- Europe
- Africa
- South America
- Asia
- Other:

How old are you? *

Your answer

What is your race? *

- White or Caucasian
- Black or African American
- American Indian or Alaska Native
- Latino or Hispanic
- Asian
- Pacific Islander or Hawaiian
- Other:

How tall are you? (in inches) *

Your answer

What is your favorite sport? *

- Soccer
- Basketball
- American Football
- Tennis
- Golf
- Baseball
- Hockey
- Rugby
- Other:

Submit