Team Bears
Tony Chen, Josh Hess, Santiago Rivera, Anand Sarathy, Kevin Thai
Professor Ethan
ISYE 4031
3 December 2019

## Executive Summary

Does better nightlife lead to better basketball performance? Does weather really make a difference to an athlete? We will demonstrate whether or not you, a sports better, should spend time and money picking and selling bets based on 14 factors outside of the basketball court. Through our analysis of 13 variables on 114 of Golden State Warriors's star shooting guard Klay Thompson, we found no relation to any of the selected external factors on his performance. The best model we came up with had just his performance related to the average rating of restaurants in the city. Our final recommendation is to discontinue research into external factors on a player's performance.

## I. Introduction

### Overview

There are many factors, both on and off the court, that can influence an NBA player's performance in a basketball game. Over an 82 game regular season (41 home and 41 away), players on 30 different teams travel more than 40,000 miles to play across 28 cities in the USA and Canada. Based on the game schedule, teams play: 4 games against the other 4 division opponents (4×4=16 games), 4 games against 6 out-of-division conference opponents (4×6=24 games), 3 games against the remaining 4 conference teams (3×4=12 games), and 2 games against teams in the opposing conference (2×15=30 games).

With the exception of a week-long All-Star break, teams play every other day, every 2 days, or on consecutive calendar days. Given these scheduling parameters and travel logistics, an individual player's performance can be impacted by various regional factors.

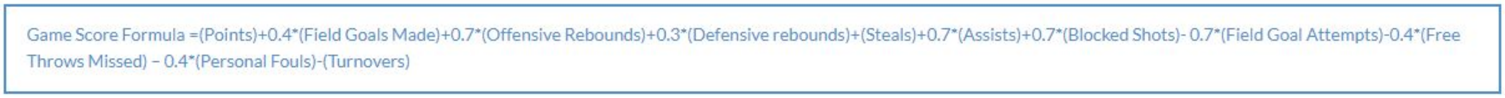
### Relevance

Sports-betting is often a high-risk high-reward endeavor. Most betting websites and advisors provide the bettor with some basic statistical insights and trends. This historical data is very important for predicting a player/team's performance, and as such there has been lots of effort put forth into analyzing traditional historical player data. Traditional data in this context are things like basic statistics (points, rebounds, assists, etc) and matchup data (how has the team/player performed against this opponent in the past).

However, there is not as much attention devoted to the regional factors and the traits of each city that can affect player performance. Because half of the games are away, players spend ample time in different environments, exposed to unfamiliar conditions. In addition, most players in the NBA are young, wealthy, and live lavish lifestyles. It is not uncommon for players to go out to nightclubs, casinos, luxurious restaurants, and other high-class establishments when playing away from home. Taking all this into account, it is possible to represent all these different off-court factors to forecast player performance. Sports betting in fact will continue to grow and has large market potential because online sports betting has recently become legal in many states. With this in mind, a lot of money can be made and continue to be made in the future if our analysis can return significant results.

## II. Problem/Goal Statement

The goal of our project is to analyze how different regional factors are related to the performance of an individual player. We will then use a regression model with these variables to forecast future performance in various cities. The metric we will use for player performance is "game score", which is a linear function of different in-game statistics such as points, assists, turnovers and so on. The full calculation for game score is as follows:



Game Score Formula =(Points)+0.4*(Field Goals Made)+0.7*(Offensive Rebounds)+0.3*(Defensive rebounds)+(Steals)+0.7*(Assists)+0.7*(Blocked Shots)- 0.7*(Field Goal Attempts)-0.4*(Free Throws Missed) – 0.4*(Personal Fouls)-(Turnovers)

As shown in the formula, positive contributions (such as Points and Field Goals Made) are positively weighted, and negative contributions (such as Turnovers and Free Throws Missed) are negatively weighted. Thus, the higher the game score the better. Regarding the observational units, we are going to use the game score for each away game in the past 3 seasons (approximately 120 games) as the dependent variable. We will classify 10+ regional factors for each city as our independent variables.

The player we are going to analyze is Klay Thompson, a 5x All-Star and 3x NBA Champion for the Golden State Warriors. He is widely regarded as the best spot-up shooter and one of the best at his position in the NBA. Because of this, Thompson generally has high game scores, and when he has a bad night, it will be easier to identify in the data. In addition, Thompson is a very durable player, rarely missing games due to injury. This will help make our model more robust when forecasting future performance using time series analysis.

Also, Klay Thompson is generally known as a fun-loving guy. On social media, you can see him going to nightclubs and having a great time in general. Obviously, you would think that some of the lifestyle choices would affect his on-court performance. The question we want to answer is
"Should you bet on Klay Thompson when he plays in 'fun' cities?"

## III. Data Description

For our data, we choose 15 different independent variables to measure the effect of various "non-traditional" factors on Klay Thompson's performance. Our 15 variables can generally be thought of in two categories: individual game data and city specific data. The individual game data are values that change for every away game Klay Thompson participated in for the previous three regular seasons (playoffs excluded). Our individual game data independent variables were:

1) Temperature
2) The Golden State Warriors win/loss streak

The temperature data for each game was calculated from https://www.wunderground.com/history**.** It should be noted that this variable refers to the average temperature that occurred on the date and in the city in which the game was played. For example, Klay Thompson played a game against the New Orleans Pelicans on 10/28/2016, so the average temperature entry for that game is 76.16 degrees fahrenheit because that is the average temperature in New Orleans on 10/28/16. These are temperature averages for away games only. We chose this variable because Klay Thompson might perform better if the temperature is neither excessively cold nor hot.

The Golden State Warriors win/loss streak was calculated from (https://www.nba.com/warriors/schedule). It should be noted that this variable refers to the streak going into the game. For example, if the warriors have won 4 in a row and are playing the Atlanta Hawks, the win/loss streak value is +4 even if they end up losing to the hawks. It should also be noted that the streak variables includes both home games and games Klay Thompson did not participate in. The logic behind this variable is that if the Warriors are on a win streak, Klay may be more relaxed and willing to enjoy the city versus if they were on a losing streak we hypothesize Klay would be more serious and less affected by his environment.

The majority of our variables are city specific data as we are mainly interested in the affects of foreign environments on Klay Thompson's game score. We make a simplifying assumption that each city specific variable has remained constant over the last three years. This assumption allows us to include several variables that we would not be able to model since historical data is unavailable. We do not anticipate this assumption significantly impacting the results of our regression analysis. The city specific variables are as follows.

1) The average casino rating
2) Average bar rating
3) Average restaurant rating
4) The city's global rating
5) The city's fun rating (in US)
6) A dummy variables for legality of marijuana
7) Number of Amusement Parks in the state/district/province the game was played
8) The city's murder rate
9) Number of building permits in the city per 10,000 people.
10) Elevation
11) Local Acceleration of Gravity

Our three average variables (casino, bar, and restaurant) were all calculated the same way. We would google "[Casinos/bars/restaurants] in [City]" and average the ratings of the top 5 results. It should be noted that we performed these searches in incognito mode to eliminate the effect of cookies and we did not filter the results. We believe the top 5 results to be sufficient as an affluent man like Klay Thompson would more than likely only go to the best/most popular locations.

The city's global rating is a measure of how innovative each city is. The data was found from https://www.innovation-cities.com/index-2019-global-city-rankings/18842/. It should be noted that we are using the innovative score, not the ranking of the city.

The city's fun rating was found from https://wallethub.com/edu/most-fun-cities-in-the-us/23455/#methodology. If a city was in the top 10 rankings than it received a fun rating of 5. If the city was in the top 30, then fun rating = 4. If the city was found in the top 50 then fun rating = 3. If the city was found in the top 100, then it received a score of 2. If the city was outside the top 100 but still appeared on the list then it received a score of 1. Cities that did not appear on the list received a score of 0.

We represented the legality of marijuana using a dummy variable where 1 = Legal for recreational use and 0 = not legal for recreational use. We used https://disa.com/map-of-marijuana-legality-by-state to find out which cities was marijuana legal for recreational use.

To find out the number of amusement parks we used https://www.ultimaterollercoaster.com/coasters/parks/states. The data is numeric and represents the amount of parks found within the state/district that Klay Thompson visited during the away game. It should be noted that this is simply a count of parks and does not measure the quality of each park.

The city's murder rate was found from https://www.statista.com/statistics/718903/murder-rate-in-us-cities-in-2015/. It is the number of reported homicides per 100,000 residents. If a city has a high murder rate, we are assuming it would be a "less fun" city. If Klay Thompson does in fact play worse in fun cities, murder rate should have an inverse relationship with game score.

To measure population in a metropolitan area, we got our data from https://www.statista.com/statistics/183600/population-of-metropolitan-areas-in-the-us/. Larger cities are often considered more vibrant, so our assumption is that a larger population leads to more fun.

Our source for building permit data was https://www.census.gov/construction/bps/txt/tb3u2018.txt. This data is useful to model how "crowded" a particular city is. To make our "crowdedness" estimate better we included another variable which accounts for population, Building Permits/10,000 residents. This variable is interesting because of Klay Thompson's known love for construction.

We used the wikipedia page of each respective away game city to find the elevation for that city. For example, Klay Thompson played an away game in Denver and so the entry for elevation corresponding to Denver is 5280 ft. because Denver's elevation is 5280 ft. Let it be known that these elevations are city-specific and not stadium-specific in which the game is played. This is most likely not an issue in collecting data since stadiums are generally closer to city centers, which is where measurements for the city elevation are usually conducted. This means that if we use elevations of the cities and not the stadiums themselves, we are still within a reasonable estimate of the precise elevation of the stadium (above sea level). We did not collect stadium-specific elevation because that data was not readily available for all of the away game stadiums, while the elevation for the cities was readily available.

We used Wolfram Alpha local acceleration of gravity utility to calculate the acceleration of gravity at each away game city. The acceleration of gravity is the constant ‘g’, which describes the acceleration of an object due to the force of gravity. We were interested in this variable because this constant describes an interaction between the Klay Thompson himself and the forces of gravity that push him down during a game. The utility is at https://www.wolframalpha.com/widgets/view.jsp?id=e856809e0d522d3153e2e7e8ec263bf2

## IV. Regression Analysis

### Variable Plots

For each of the independent variables we are examining, the variable plots were created by using the plot function for all the continuous variables, and using the boxplot function for the few discrete variables. We plotted each of the independent variables on the x-axis and the resulting game scores on the y-axis. From these variable plots shown in the appendix, there are some preliminary findings that are not very surprising when attempting to find trends and relationships in each of the independent variables.

For example, we can see that in games where the Golden State Warriors win, there is a noticeably higher average game score compared to when  losses which should come as no surprise. Another relationship that is not surprising is that Klay Thompson’s game scores are higher on average when the Golden State Warriors are on a win streak. This is related to the previous binary relationship of win / loss, but indicates a stronger relationship overall to support the statement that Klay Thompson has a higher average game score in wins compared to losses.

Some notable findings that are not as obvious as winning and losing include: average restaurant rating and legalization of marijuana. With the exception of a few outliers, the restaurant vs game score plot appears to show a moderate positive linear relationship as the average restaurant ratings increase, so does Klay Thompson's game score. When looking at states that have not legalized marijuana, Klay Thompson's average game score is a few points higher than states that have legalized marijuana. The 25th and 75th percentiles of average game score are higher in states that have not legalized weed compared to the states that have legalized weed. It appears to be evident that Klay Thompson's average game score is noticeably better in states that have not legalized weed.

The remaining independent variables do not appear to show any significant trends in their individual plots against game score. We will continue to run further analysis to see if we are able to find any other noticeable trends.

**Full Model**

Figure 3 shows a summary of the full model. It is immediately obvious that, since the p value of each of the variables is very high (the minimum is 0.362), we have low confidence that any of the variables are correlated with Klay Thompson's game score (low confidence that β != 0). Additionally, the adjusted R-squared value is very low (-0.08851) indicating very high error in our fitted model. Finally, the f-statistic is very high (0.2932) which indicates low confidence that any of the variables are correlated with game score.

Figure 1 shows that the mean VIF of the variables is greater than 1 (~1.65). This suggests severe multicollinearity between the variables. However, since none of the VIFs of any variables is greater than 10, it isn't immediately clear which ones to remove.

Figure 2 shows that the correlation of any pair of variables is very low with the greatest being ~0.39 between number of amusement parks and acceleration of gravity. This indicates low correlation between a pair of variables. However, this correlation is not meaningless and something that should be considered if we're looking for parsimony between models with both variables.

**Model Selection**

For our model selection, we looked at Adjusted R-Squared, $C_p$ value, Residual Sum of Squares, and AIC value. Our R outputs can be viewed in Figure 4. From the output, we observe that the best model is one where the only independent variable is Restaurants. We can tell because this model has the lowest $C_p$ and AIC values, the greatest Adjusted R-Squared value, and the lowest Residual Sum of Squares.

The second best model which was close in those metrics was a model that only included Restaurants and Average Temperature. Each subsequent worse model from regsubset only served to add a variable to the previous worst model. This indicates that our independent variables are overall poor predictors of game score and worsen the model as they are added.

**Game Score vs Restaurants Model**

Figure 5 shows the summary for the Game Score vs Restaurants Model. The p values for the coefficients are once again high (0.287 for Restaurants), indicating low-medium confidence that an increase in Restaurants has any effect on Klay Thompson's Game Score. Additionally, the R-squared values are very low (0.001291 for adjusted) indicating that not much of the error from the mean is captured by the model. Figure 6 shows the scatterplot and fitted line for this model. The line has a slight positive slope but it is clear that it doesn't track the points very well. Since this model only has one independent variable, there is no multicollinearity to look at.

**Game Score vs Restaurants + Average Temperature**

Figure 7 shows the summary for the Game Score vs Restaurants + Average Temperature model (our second best model). The p-values are still high (0.247 for Restaurants, 0.410 for Average Temperature which are both much greater than an alpha of 0.1) which indicates low-medium confidence that an increase in one of these variables has an effect on Klay Thompson's Game Score. Similar to the model with only Restaurants as an independent variable, the r-squared values are very low and the p-value of the f-statistic is high. Figure 8 shows the variance inflation factor (VIF) of Restaurants and Average Temperature. While the mean is above 1, it is just barely so. The figure also shows that the correlation between the variables is low (magnitude of 0.123) which suggests that the two variables don't influence each other drastically.

**Outliers**

The leverage values for each model are shown in Figure 9. In model 1, this indicates that our 31st and 60th data points are outliers with respect to their x values.

The studentized deleted residuals for each model that exceed the t-value limit are shown in Figure 10. We can see that games 81 and 98 are outliers with respect to the Game Score for each model.

The Cook's distance of each model is shown in Figure 11. From the analysis, we can see that none of the observations should be considered influential. This means that removing an observation will not have a significant effect on our fit line.

**Residual Analysis**

We need to conduct residual analysis on our model because that way we can use our error terms to add another layer of security to us that our model is valid and can be used to make recommendations to our client. This means we need to check the validity of the regression assumptions. To do this, we look at the residuals of our model through the residual plots (see appendix residual plot figures) to see if represent the constant variance assumption and correct functional form with residuals spread out randomly. The code to gather each of these plots was:

```
> plot(<indep. variable>, residuals(data_LM), pch=16,col="blue", ylab="e or Residuals")
```

From the residual plots, we can see that Bars independent variables seem to **not** fit these regression assumptions:

1. Constant Variance
2. Correct functional form (Residuals spread out randomly)

Upon further inspection, we noted that the residual plot for this independent variable looked like it fanned upward because of two outliers towards the end of the x axis. Without these, and with more data for bars with low ratings, we would likely see the residuals not fan out and have constant variance.

For further residual analysis, we need to look at the normality plots and tests for our model. According to the normality probability plot, if the points fall along a straight line, then the normality assumption is checked. You can check the normality probability plots for the residuals by looking in the appendix and also the code to implement those plots and download the nortest suite in R. From the normal probability plots, we can see that all independent variables seem to follow the line pretty closely, especially for the bulk of points. Just to be sure that we are not violating normality assumptions, we ran anderson darling tests for our model regarding each independent variable. The code for these tests and corresponding output can be found in the appendix. From these Anderson Darling tests we found that they all passed and that the normality assumption was checked when setting our alpha in this test to .01, .05, and even .1. Now we can be more confident that we have a valid model and that no transformations are needed for the model.

**V. Conclusion**

“Should you bet on Klay Thompson when he plays in ‘fun’ cities?”

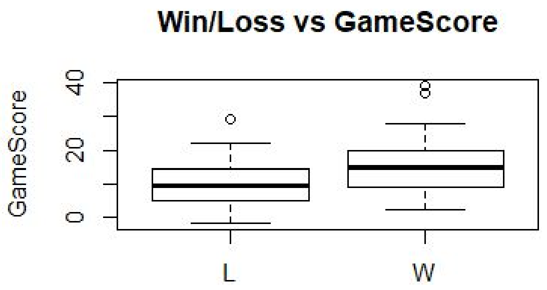
Our regression analysis on Klay Thompson’s game score indicates that the “fun” factor of different cities does not have a significant effect on his performance. None of the variables we selected had a significant relationship to game score, and from this we draw the conclusion that these external factors outside of the court have little effect on performance.

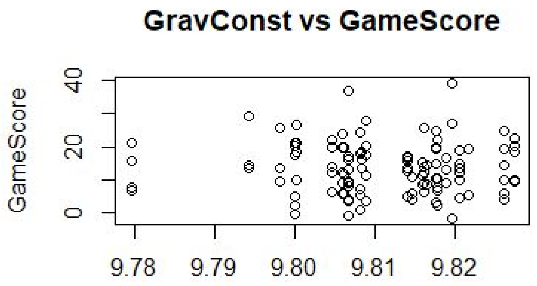
Our recommendation is to discontinue research into “non-traditional” variables impacting game performance, as there was scant evidence of a possibility of a relationship.

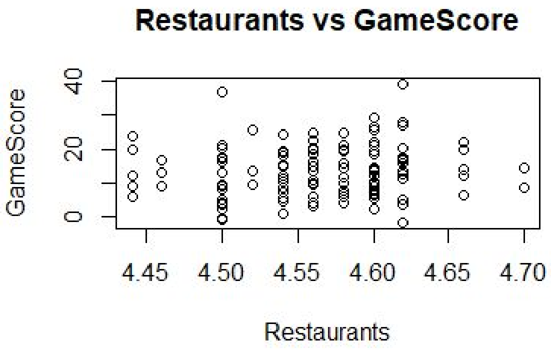
One explanation for the lack of any relationship could be the use of sports psychologists whose purpose is to help players maintain performance in game by keeping a strong mental state in spite of anything going on outside of the game.

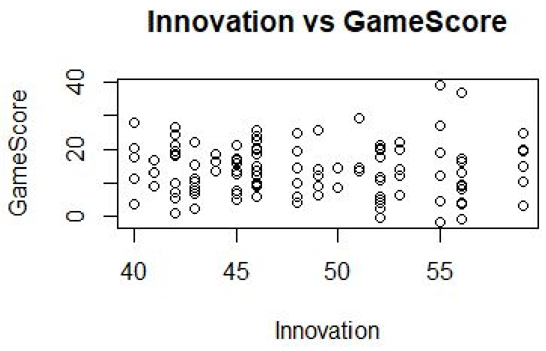
## VI. Appendix

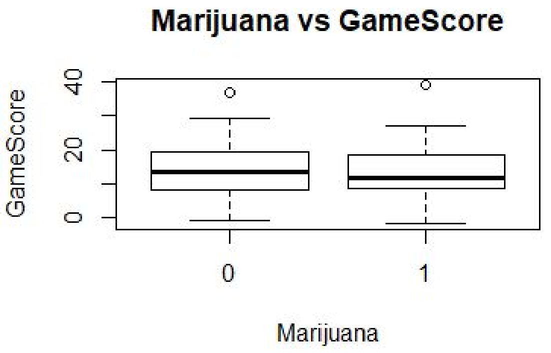
### Variable Plots:

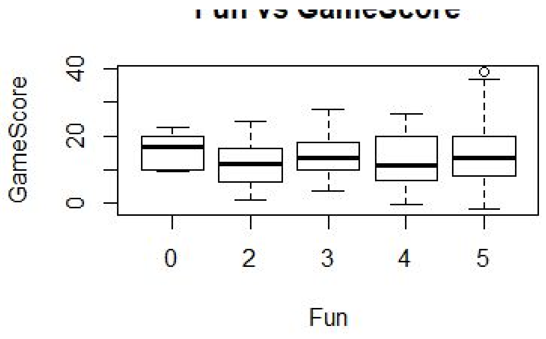

Win/Loss vs GameScore

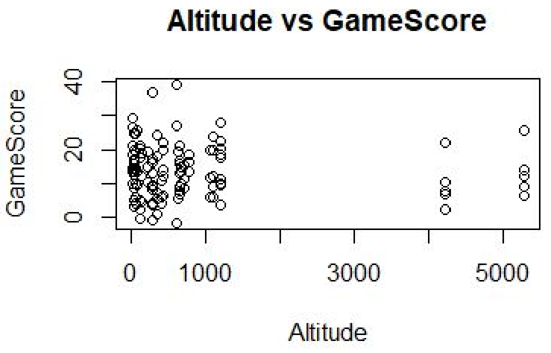

GravConst vs GameScore

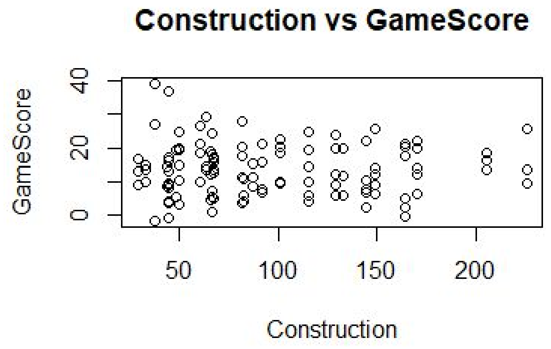

Restaurants vs GameScore

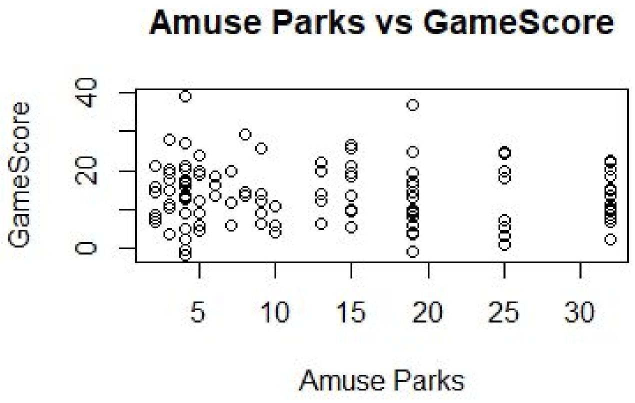

Innovation vs GameScore

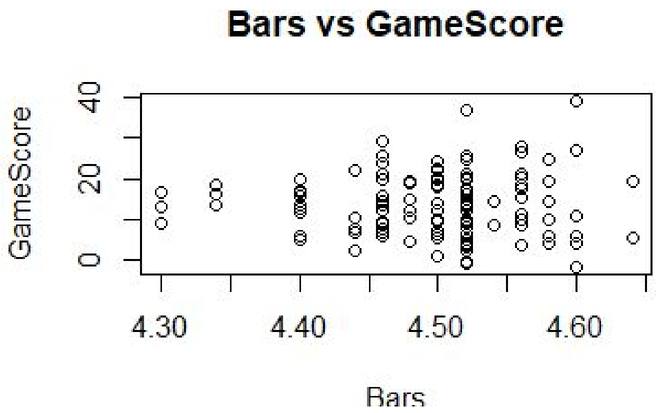

Marijuana vs GameScore

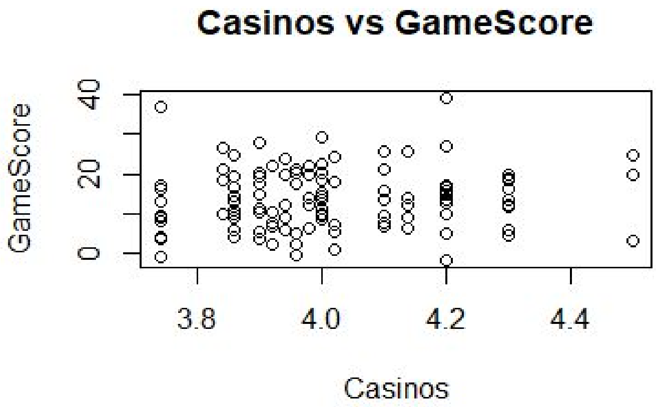

Fun vs GameScore

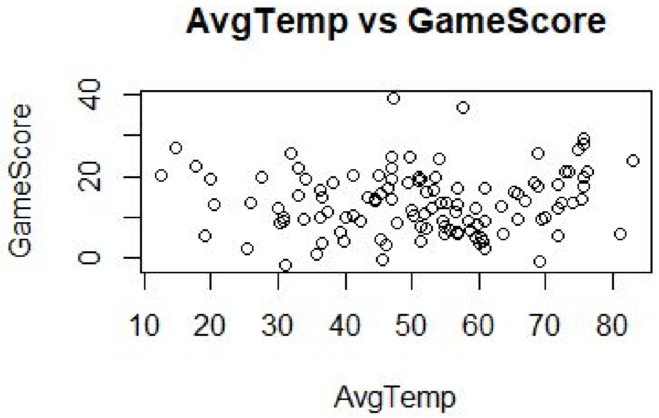

Altitude vs GameScore

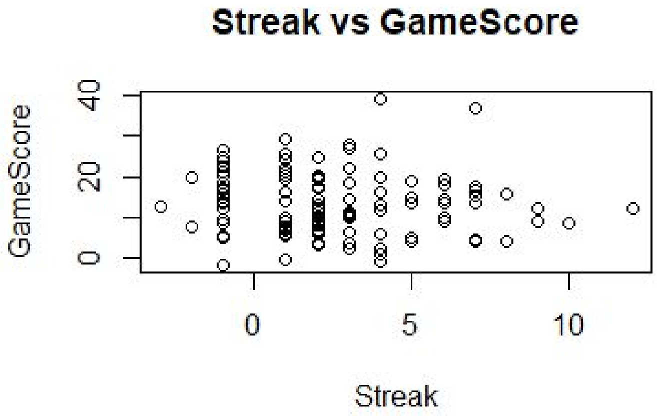

Construction vs GameScore

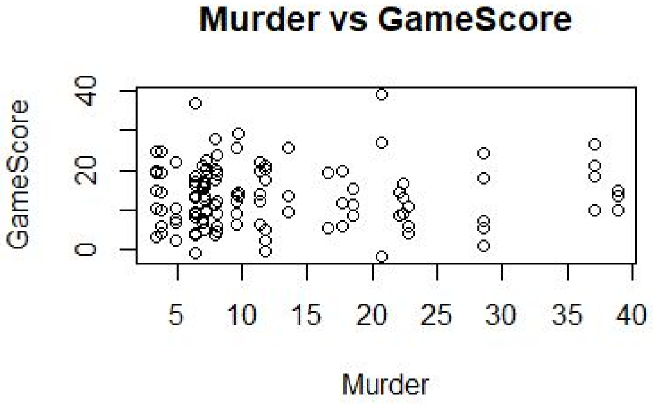

Amuse Parks vs GameScore
GameScore
0
20
40
5
10
15
20
25
30
Amuse Parks

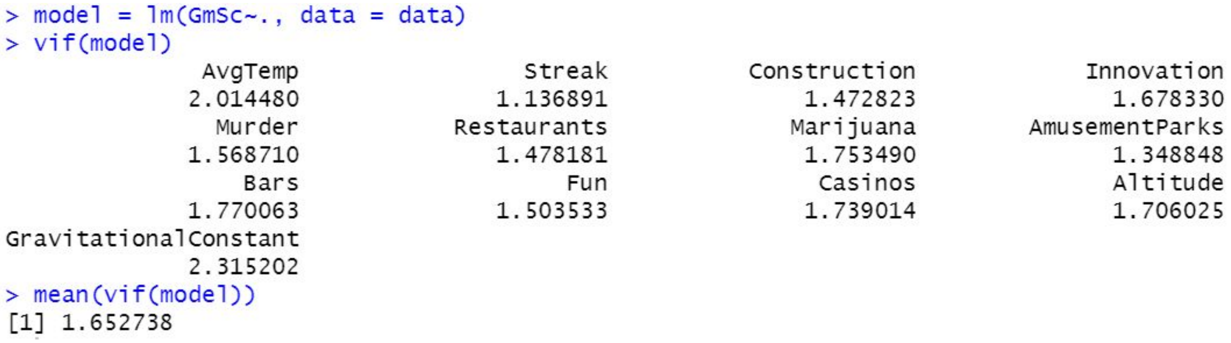

Bars vs GameScore
GameScore
0
20
40
4.30
4.40
4.50
4.60
Bars


Casinos vs GameScore
GameScore
0
20
40
3.8
4.0
4.2
4.4
Casinos


AvgTemp vs GameScore
GameScore
0
20
40
10
20
30
40
50
60
70
80
AvgTemp

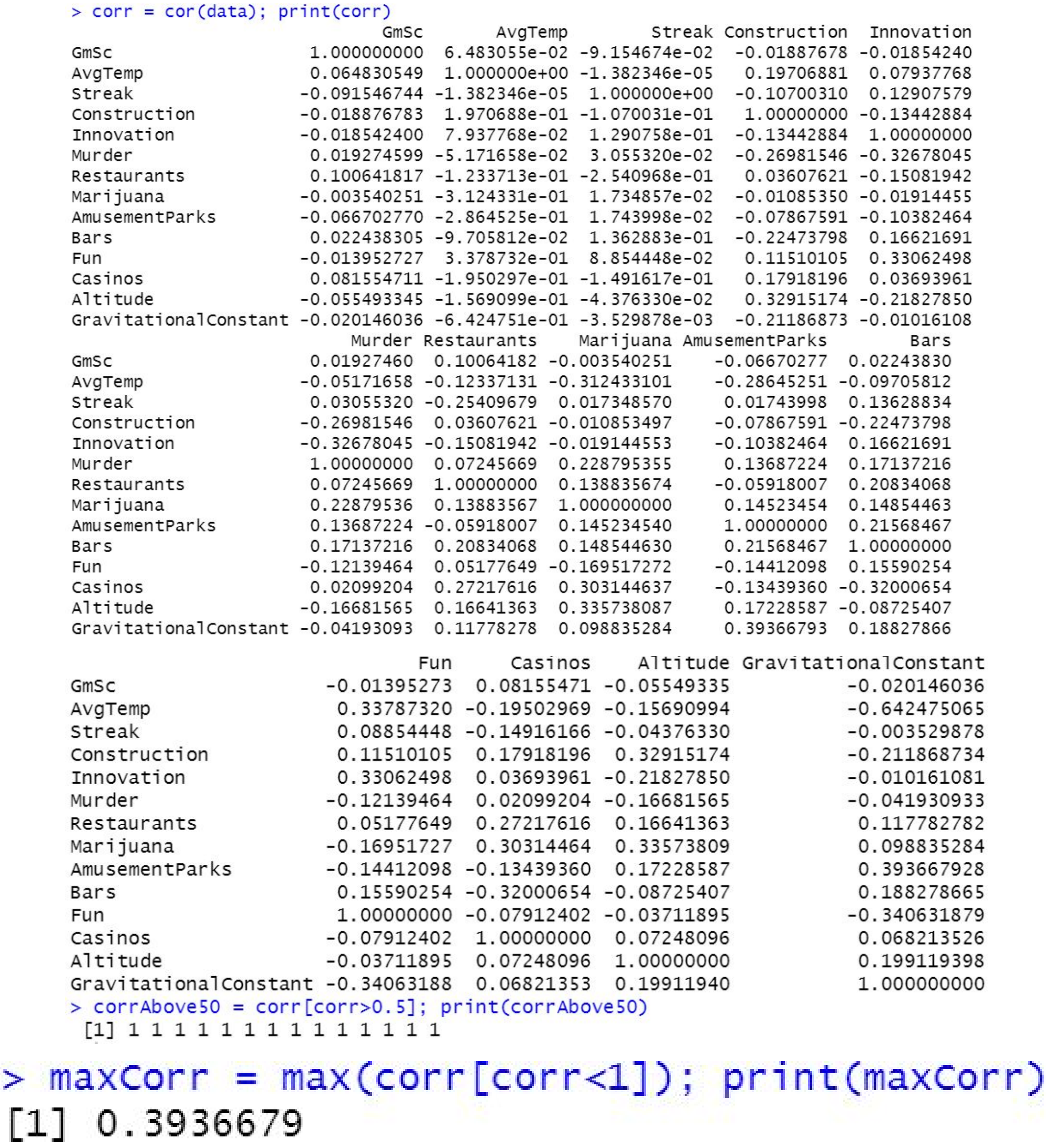

Streak vs GameScore
GameScore
0
20
40
0
5
10
Streak


Murder vs GameScore
GameScore
0
20
40
5
10
15
20
25
30
35
40
Murder



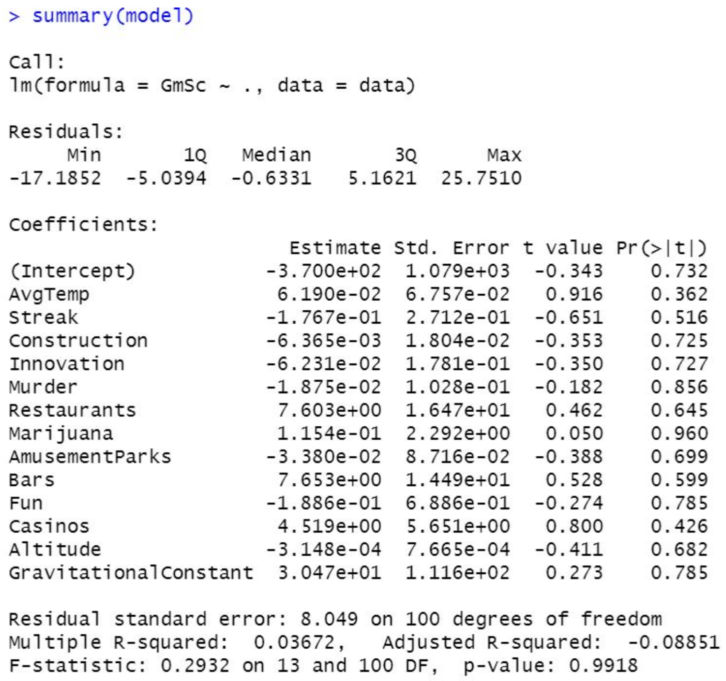
```
> model = lm(GmSc~., data = data)
> vif(model)
              AvgTemp                Streak          Construction            Innovation 
             2.014480              1.136891              1.472823              1.678330 
               Murder           Restaurants             Marijuana        AmusementParks 
             1.568710              1.478181              1.753490              1.348848 
                 Bars                   Fun               Casinos              Altitude 
             1.770063              1.503533              1.739014              1.706025 
GravitationalConstant 
             2.315202 
> mean(vif(model))
[1] 1.652738
```

**Figure 1:  VIF and mean VIF for the variables.**



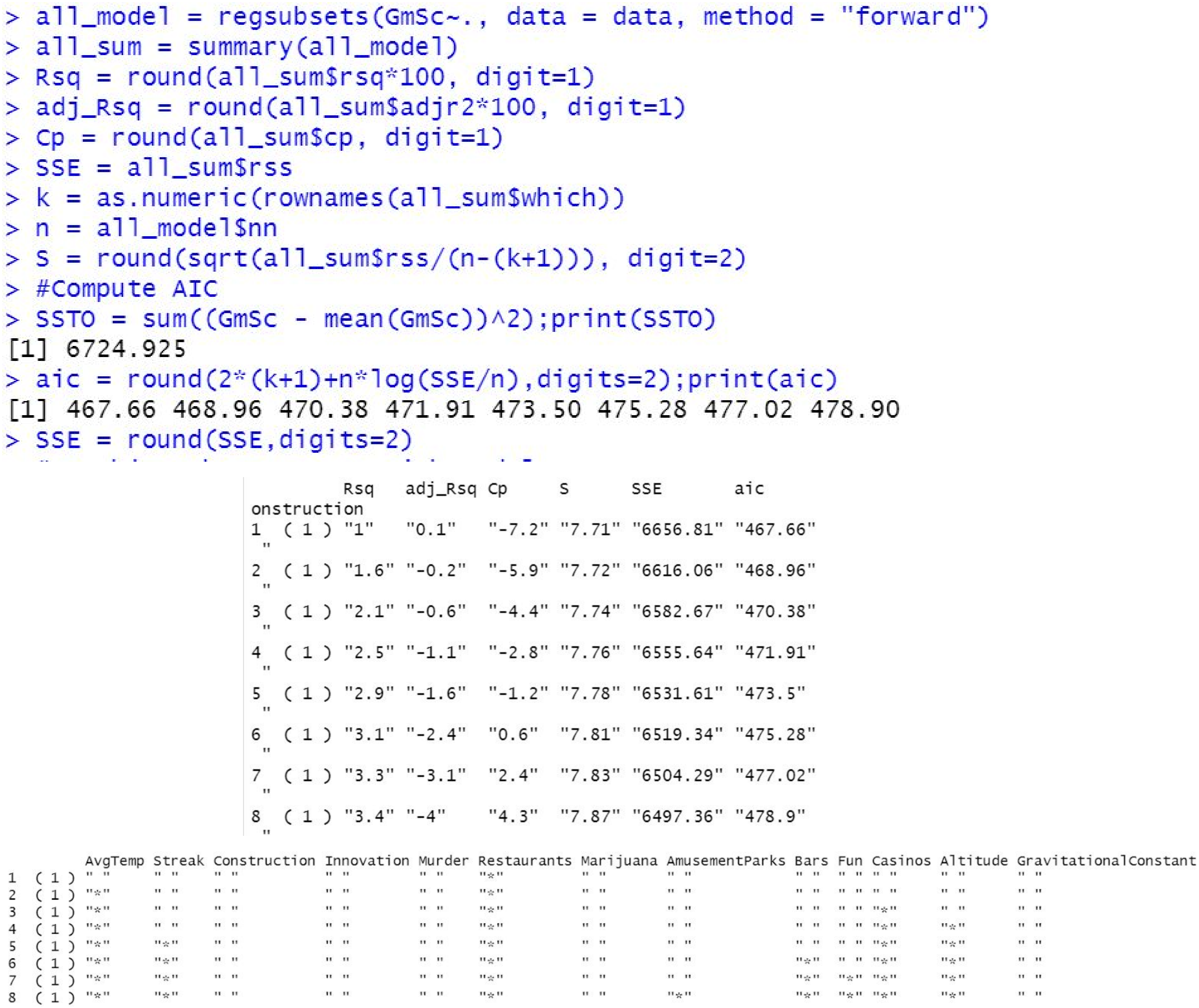
```
> corr = cor(data); print(corr)
                             GmSc      AvgTemp        Streak Construction  Innovation
GmSc                  1.000000000  6.483055e-02 -9.154674e-02  -0.01887678 -0.01854240
AvgTemp               0.064830549  1.000000e+00 -1.382346e-05   0.19706881  0.07937768
Streak               -0.091546744 -1.382346e-05  1.000000e+00  -0.10700310  0.12907579
Construction         -0.018876783  1.970688e-01 -1.070031e-01   1.00000000 -0.13442884
Innovation           -0.018542400  7.937768e-02  1.290758e-01  -0.13442884  1.00000000
Murder                0.019274599 -5.171658e-02  3.055320e-02  -0.26981546 -0.32678045
Restaurants           0.100641817 -1.233713e-01 -2.540968e-01   0.03607621 -0.15081942
Marijuana            -0.003540251 -3.124331e-01  1.734857e-02  -0.01085350 -0.01914455
AmusementParks       -0.066702770 -2.864525e-01  1.743998e-02  -0.07867591 -0.10382464
Bars                  0.022438305 -9.705812e-02  1.362883e-01  -0.22473798  0.16621691
Fun                  -0.013952727  3.378732e-01  8.854448e-02   0.11510105  0.33062498
Casinos               0.081554711 -1.950297e-01 -1.491617e-01   0.17918196  0.03693961
Altitude             -0.055493345 -1.569099e-01 -4.376330e-02   0.32915174 -0.21827850
GravitationalConstant -0.020146036 -6.424751e-01 -3.529878e-03  -0.21186873 -0.01016108
                          Murder Restaurants    Marijuana AmusementParks        Bars
GmSc                  0.01927460  0.10064182 -0.003540251    -0.06670277  0.02243830
AvgTemp              -0.05171658 -0.12337131 -0.312433101    -0.28645251 -0.09705812
Streak                0.03055320 -0.25409679  0.017348570     0.01743998  0.13628834
Construction         -0.26981546  0.03607621 -0.010853497    -0.07867591 -0.22473798
Innovation           -0.32678045 -0.15081942 -0.019144553    -0.10382464  0.16621691
Murder                1.00000000  0.07245669  0.228795355     0.13687224  0.17137216
Restaurants           0.07245669  1.00000000  0.138835674    -0.05918007  0.20834068
Marijuana             0.22879536  0.13883567  1.000000000     0.14523454  0.14854463
AmusementParks        0.13687224 -0.05918007  0.145234540     1.00000000  0.21568467
Bars                  0.17137216  0.20834068  0.148544630     0.21568467  1.00000000
Fun                  -0.12139464  0.05177649 -0.169517272    -0.14412098  0.15590254
Casinos               0.02099204  0.27217616  0.303144637    -0.13439360 -0.32000654
Altitude             -0.16681565  0.16641363  0.335738087     0.17228587 -0.08725407
GravitationalConstant -0.04193093  0.11778278  0.098835284     0.39366793  0.18827866

                             Fun     Casinos    Altitude GravitationalConstant
GmSc                 -0.01395273  0.08155471 -0.05549335          -0.020146036
AvgTemp               0.33787320 -0.19502969 -0.15690994          -0.642475065
Streak                0.08854448 -0.14916166 -0.04376330          -0.003529878
Construction          0.11510105  0.17918196  0.32915174          -0.211868734
Innovation            0.33062498  0.03693961 -0.21827850          -0.010161081
Murder               -0.12139464  0.02099204 -0.16681565          -0.041930933
Restaurants           0.05177649  0.27217616  0.16641363           0.117782782
Marijuana            -0.16951727  0.30314464  0.33573809           0.098835284
AmusementParks       -0.14412098 -0.13439360  0.17228587           0.393667928
Bars                  0.15590254 -0.32000654 -0.08725407           0.188278665
Fun                   1.00000000 -0.07912402 -0.03711895          -0.340631879
Casinos              -0.07912402  1.00000000  0.07248096           0.068213526
Altitude             -0.03711895  0.07248096  1.00000000           0.199119398
GravitationalConstant -0.34063188  0.06821353  0.19911940           1.000000000
> corrAbove50 = corr[corr>0.5]; print(corrAbove50)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1
> maxCorr = max(corr[corr<1]); print(maxCorr)
[1] 0.3936679
```

**Figure 2: Correlation information for the full model.**



```
> summary(model)

Call:
lm(formula = GmSc ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.1852  -5.0394  -0.6331   5.1621  25.7510 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -3.700e+02  1.079e+03  -0.343    0.732
AvgTemp                6.190e-02  6.757e-02   0.916    0.362
Streak                -1.767e-01  2.712e-01  -0.651    0.516
Construction          -6.365e-03  1.804e-02  -0.353    0.725
Innovation            -6.231e-02  1.781e-01  -0.350    0.727
Murder                -1.875e-02  1.028e-01  -0.182    0.856
Restaurants            7.603e+00  1.647e+01   0.462    0.645
Marijuana              1.154e-01  2.292e+00   0.050    0.960
AmusementParks        -3.380e-02  8.716e-02  -0.388    0.699
Bars                   7.653e+00  1.449e+01   0.528    0.599
Fun                   -1.886e-01  6.886e-01  -0.274    0.785
Casinos                4.519e+00  5.651e+00   0.800    0.426
Altitude              -3.148e-04  7.665e-04  -0.411    0.682
GravitationalConstant  3.047e+01  1.116e+02   0.273    0.785

Residual standard error: 8.049 on 100 degrees of freedom
Multiple R-squared:  0.03672,	Adjusted R-squared:  -0.08851 
F-statistic: 0.2932 on 13 and 100 DF,  p-value: 0.9918
```

**Figure 3: Summary of full model.**

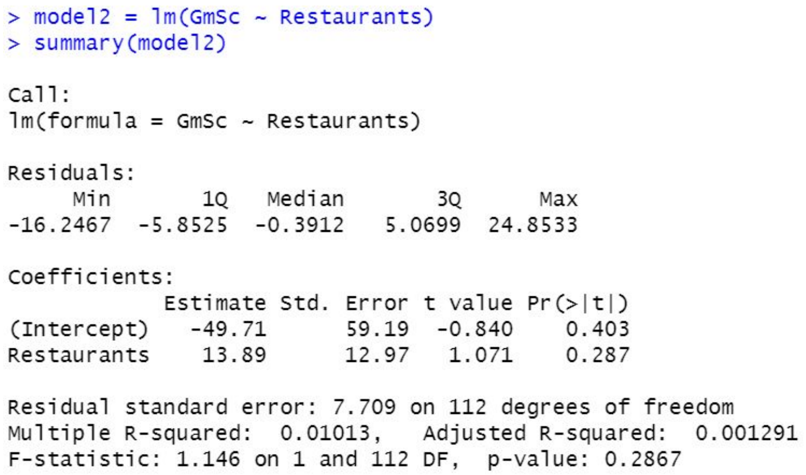


```
> all_model = regsubsets(GmSc~., data = data, method = "forward")
> all_sum = summary(all_model)
> Rsq = round(all_sum$rsq*100, digit=1)
> adj_Rsq = round(all_sum$adjr2*100, digit=1)
> Cp = round(all_sum$cp, digit=1)
> SSE = all_sum$rss
> k = as.numeric(rownames(all_sum$which))
> n = all_model$nn
> S = round(sqrt(all_sum$rss/(n-(k+1))), digit=2)
> #Compute AIC
> SSTO = sum((GmSc - mean(GmSc))^2);print(SSTO)
[1] 6724.925
> aic = round(2*(k+1)+n*log(SSE/n),digits=2);print(aic)
[1] 467.66 468.96 470.38 471.91 473.50 475.28 477.02 478.90
> SSE = round(SSE,digits=2)
```

```
           Rsq   adj_Rsq Cp     S      SSE       aic
onstruction
1  ( 1 ) "1"   "0.1"   "-7.2" "7.71" "6656.81" "467.66"
 "
2  ( 1 ) "1.6" "-0.2"  "-5.9" "7.72" "6616.06" "468.96"
 "
3  ( 1 ) "2.1" "-0.6"  "-4.4" "7.74" "6582.67" "470.38"
 "
4  ( 1 ) "2.5" "-1.1"  "-2.8" "7.76" "6555.64" "471.91"
 "
5  ( 1 ) "2.9" "-1.6"  "-1.2" "7.78" "6531.61" "473.5"
 "
6  ( 1 ) "3.1" "-2.4"  "0.6"  "7.81" "6519.34" "475.28"
 "
7  ( 1 ) "3.3" "-3.1"  "2.4"  "7.83" "6504.29" "477.02"
 "
8  ( 1 ) "3.4" "-4"    "4.3"  "7.87" "6497.36" "478.9"
 "
```

```
         AvgTemp Streak Construction Innovation Murder Restaurants Marijuana AmusementParks Bars Fun Casinos Altitude GravitationalConstant
1  ( 1 ) " "     " "    " "          " "        " "    "*"         " "       " "            " "  " " " "     " "      " "
2  ( 1 ) "*"     " "    " "          " "        " "    "*"         " "       " "            " "  " " " "     " "      " "
3  ( 1 ) "*"     " "    " "          " "        " "    "*"         " "       " "            " "  " " "*"     " "      " "
4  ( 1 ) "*"     " "    " "          " "        " "    "*"         " "       " "            " "  " " "*"     "*"      " "
5  ( 1 ) "*"     "*"    " "          " "        " "    "*"         " "       " "            " "  " " "*"     "*"      " "
6  ( 1 ) "*"     "*"    " "          " "        " "    "*"         " "       " "            "*"  " " "*"     "*"      " "
7  ( 1 ) "*"     "*"    " "          " "        " "    "*"         " "       " "            "*"  "*" "*"     "*"      " "
8  ( 1 ) "*"     "*"    " "          " "        " "    "*"         " "       "*"            "*"  "*" "*"     "*"      " "
```

**Figure 4: Model Selection using indicators.**



```
> model2 = lm(GmSc ~ Restaurants)
> summary(model2)

Call:
lm(formula = GmSc ~ Restaurants)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.2467  -5.8525  -0.3912   5.0699  24.8533 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -49.71      59.19  -0.840    0.403
Restaurants    13.89      12.97   1.071    0.287

Residual standard error: 7.709 on 112 degrees of freedom
Multiple R-squared:  0.01013,	Adjusted R-squared:  0.001291 
F-statistic: 1.146 on 1 and 112 DF,  p-value: 0.2867
```

**Figure 5: Summary of Game Score vs Restaurants Model**

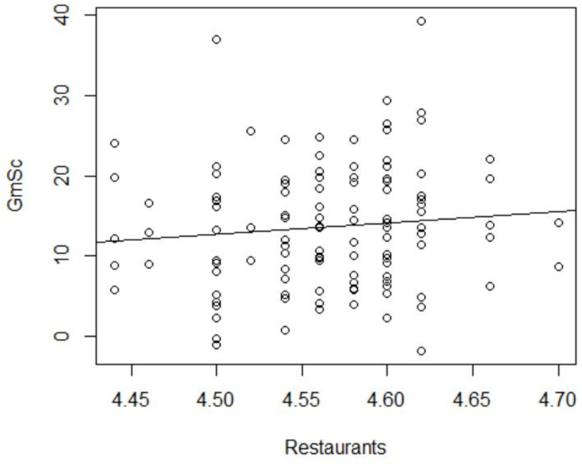
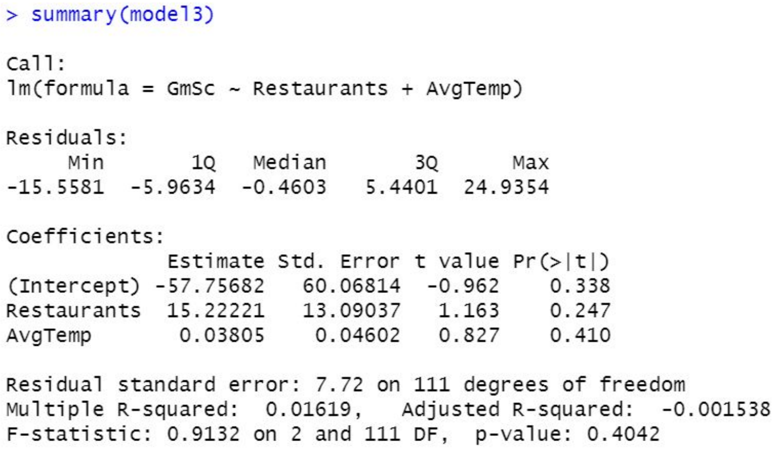
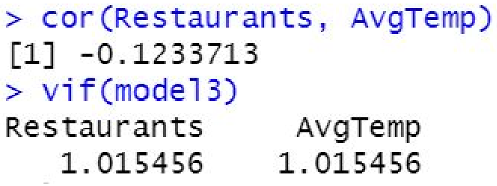


**Figure 6: Graph of Game Score vs Restaurants Model**



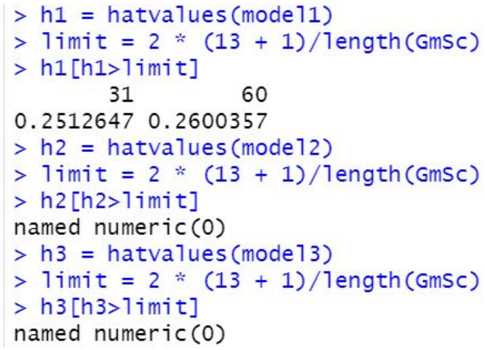
```
> summary(model3)

Call:
lm(formula = GmSc ~ Restaurants + AvgTemp)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.5581  -5.9634  -0.4603   5.4401  24.9354 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.75682   60.06814  -0.962    0.338
Restaurants  15.22221   13.09037   1.163    0.247
AvgTemp       0.03805    0.04602   0.827    0.410

Residual standard error: 7.72 on 111 degrees of freedom
Multiple R-squared:  0.01619,	Adjusted R-squared:  -0.001538 
F-statistic: 0.9132 on 2 and 111 DF,  p-value: 0.4042
```

**Figure 7: Summary of Game Score vs Restaurants + Average Temperature Model**



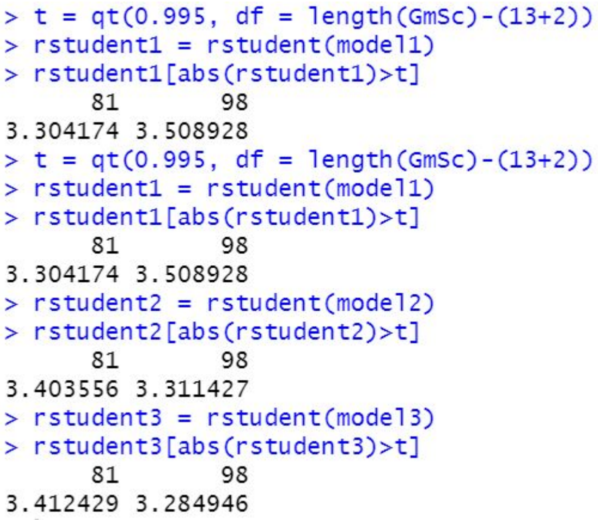
```
> cor(Restaurants, AvgTemp)
[1] -0.1233713
> vif(model3)
Restaurants     AvgTemp 
   1.015456    1.015456 
```

**Figure 8: Multicollinearity information for Game Score vs Restaurants + Average Temperature Model**



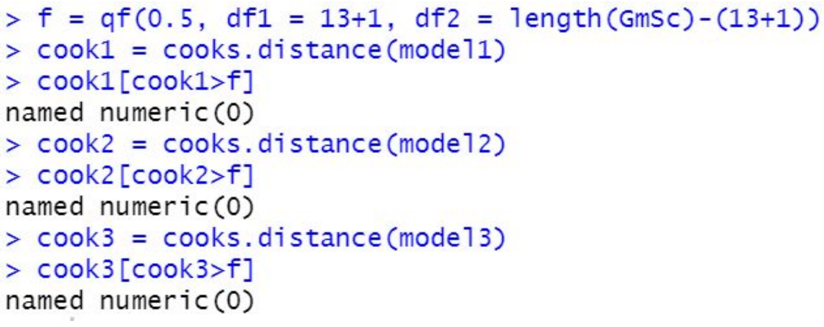
```
> h1 = hatvalues(model1)
> limit = 2 * (13 + 1)/length(GmSc)
> h1[h1>limit]
       31        60
0.2512647 0.2600357
> h2 = hatvalues(model2)
> limit = 2 * (13 + 1)/length(GmSc)
> h2[h2>limit]
named numeric(0)
> h3 = hatvalues(model3)
> limit = 2 * (13 + 1)/length(GmSc)
> h3[h3>limit]
named numeric(0)
```

**Figure 9: Hat Values for each model.**



```
> t = qt(0.995, df = length(GmSc)-(13+2))
> rstudent1 = rstudent(model1)
> rstudent1[abs(rstudent1)>t]
      81       98
3.304174 3.508928
> t = qt(0.995, df = length(GmSc)-(13+2))
> rstudent1 = rstudent(model1)
> rstudent1[abs(rstudent1)>t]
      81       98
3.304174 3.508928
> rstudent2 = rstudent(model2)
> rstudent2[abs(rstudent2)>t]
      81       98
3.403556 3.311427
> rstudent3 = rstudent(model3)
> rstudent3[abs(rstudent3)>t]
      81       98
3.412429 3.284946
```

**Figure 10: Analysis of Outliers with respect to Y values.**

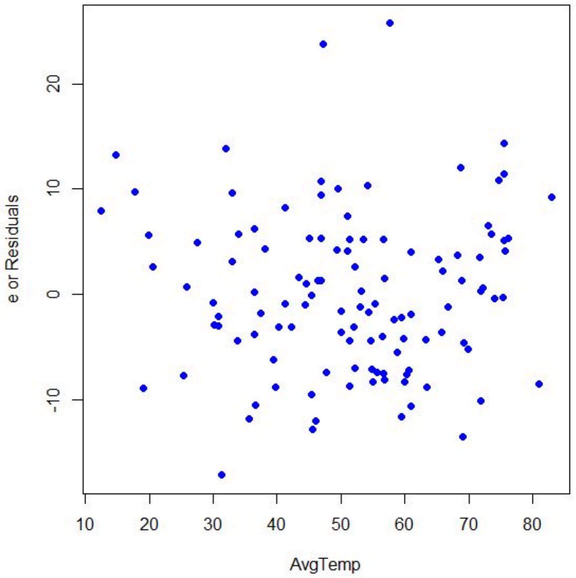


```
> f = qf(0.5, df1 = 13+1, df2 = length(GmSc)-(13+1))
> cook1 = cooks.distance(model1)
> cook1[cook1>f]
named numeric(0)
> cook2 = cooks.distance(model2)
> cook2[cook2>f]
named numeric(0)
> cook3 = cooks.distance(model3)
> cook3[cook3>f]
named numeric(0)
```

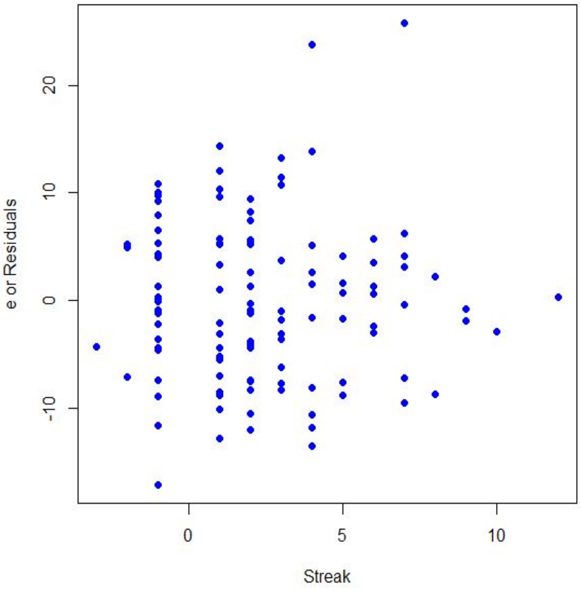
**Figure 11: Analysis of Influential Points.**

**Residual Plots:**

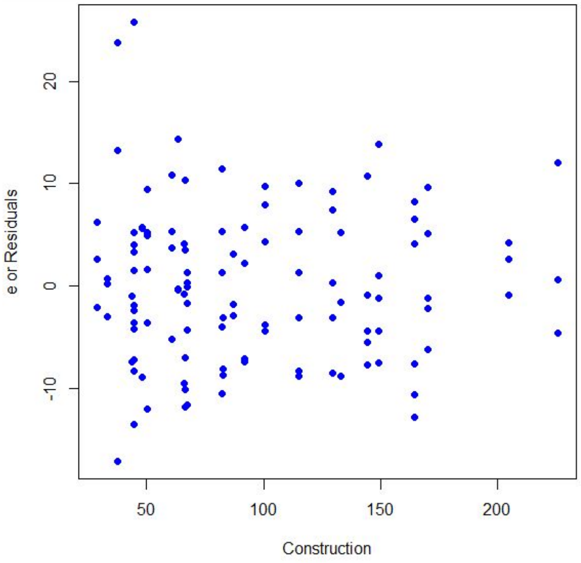
Average Temperature Residuals



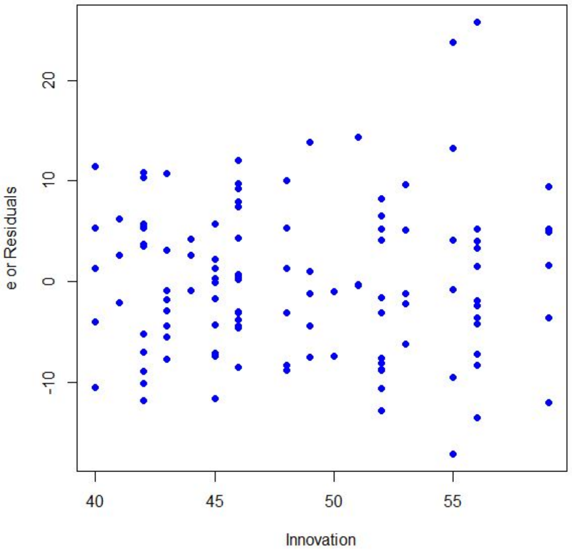
Streak Residuals



Construction Residuals

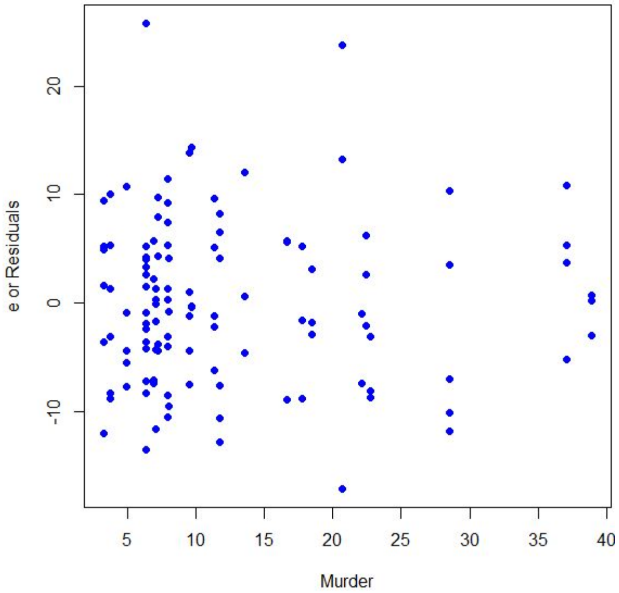


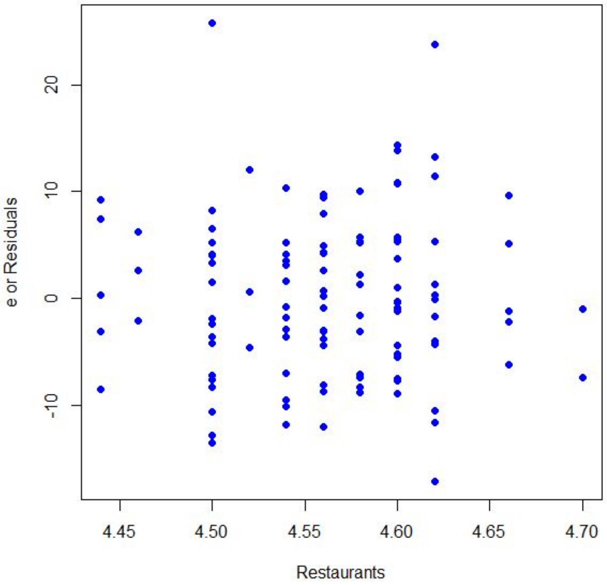
Innovation Residuals



Murder Residuals

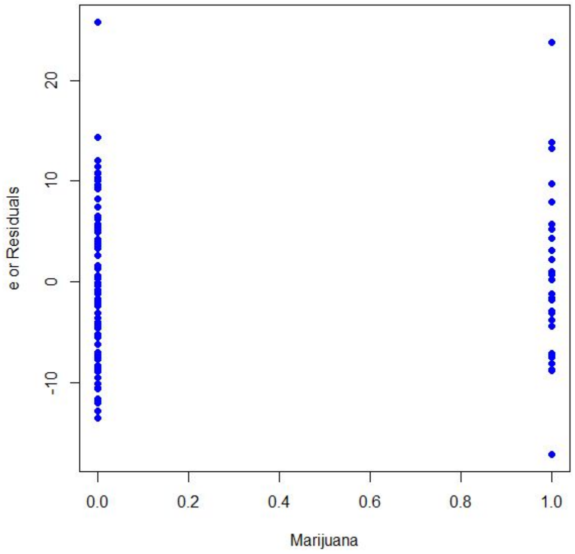
Restaurants Residuals

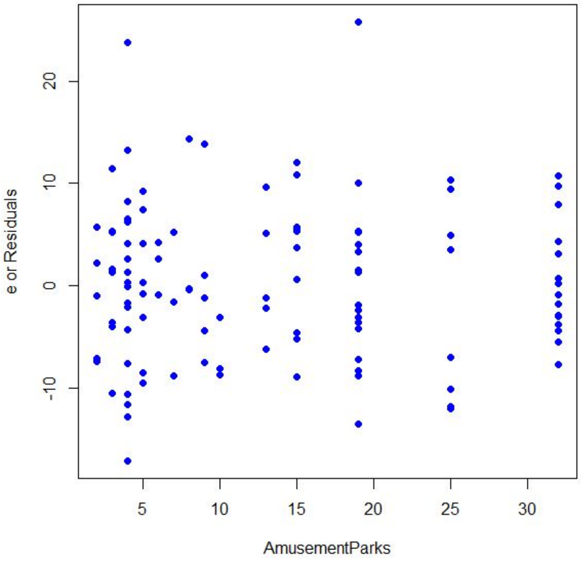




Marijuana Residuals

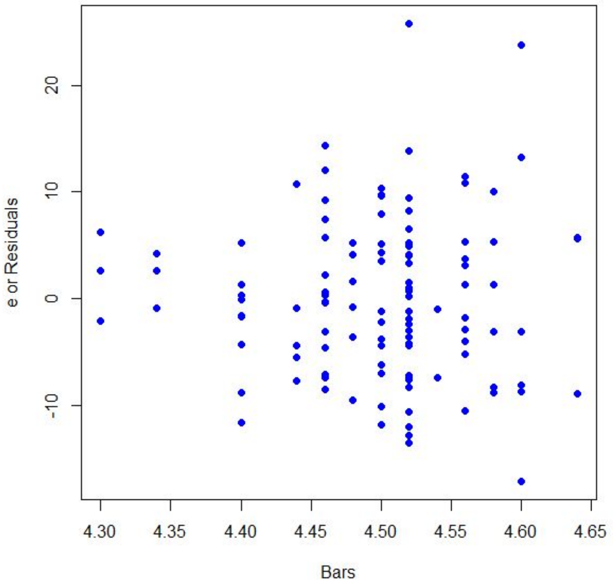
Amusement Parks Residuals

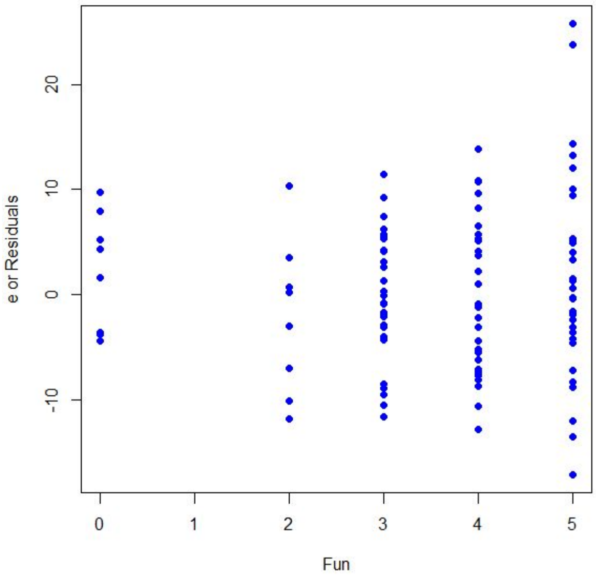




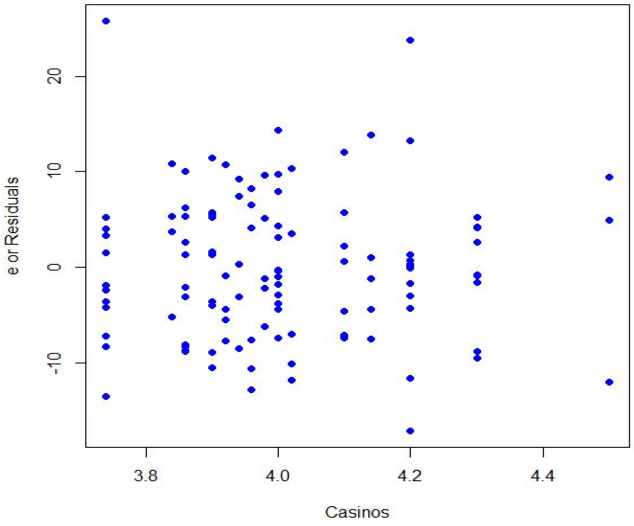
Bars Residuals

Fun Residuals

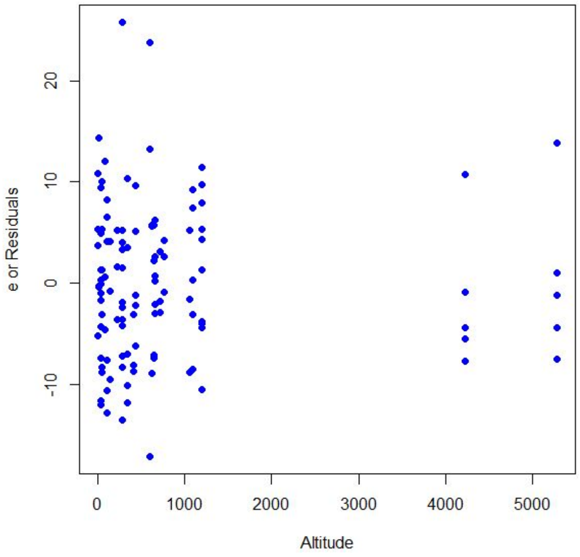




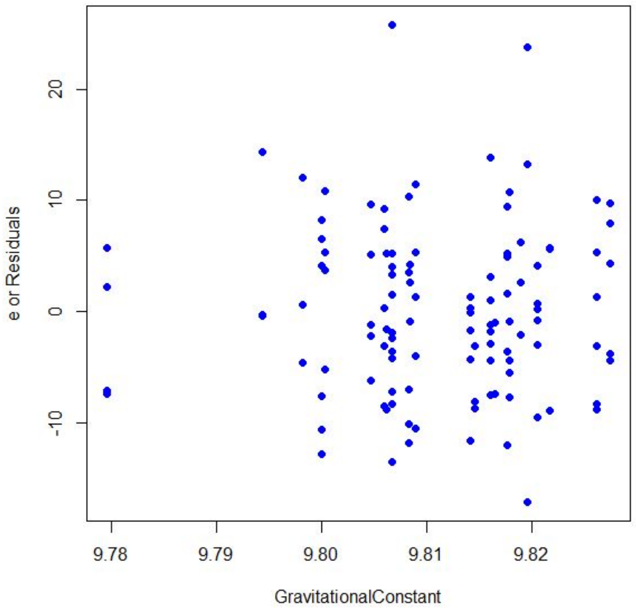
Casinos Residuals



Altitude Residuals



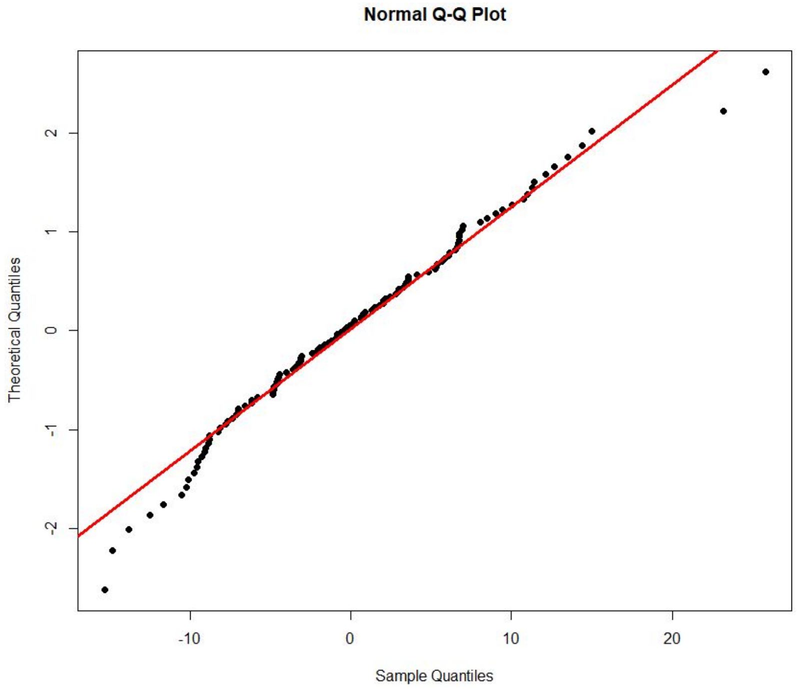
## Acceleration of Gravity Residuals



**Normal Probability Plots**

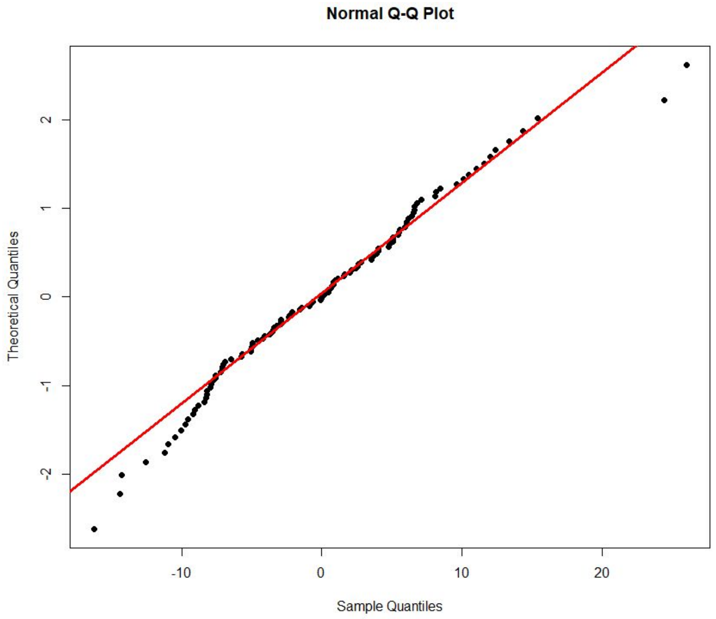
For Average Temperature:

```
> qqnorm(resid(lm(GmSc ~ AvgTemp)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ AvgTemp)), datax=TRUE, lwd=3, col="red")
```



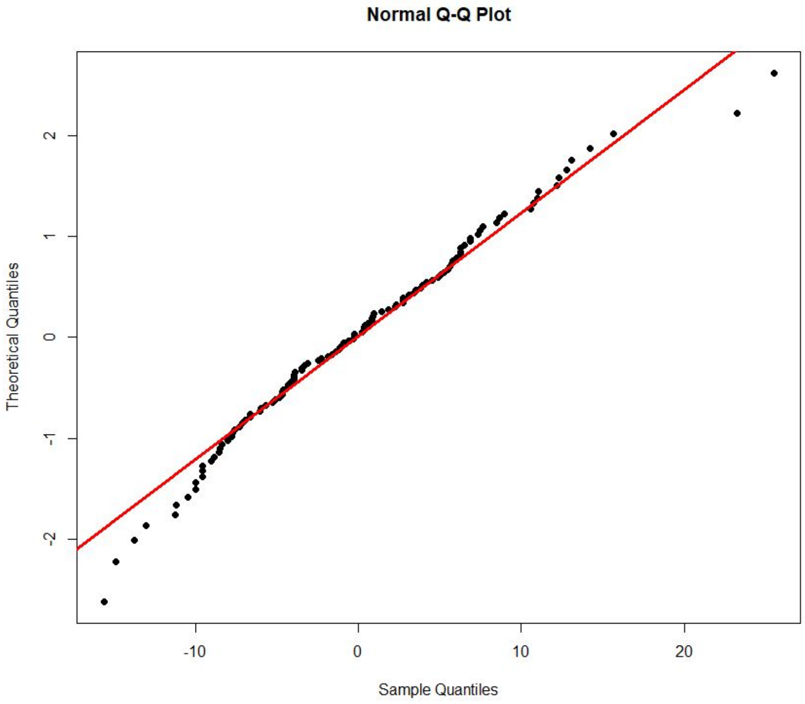
For Streak:

```
> qqnorm(resid(lm(GmSc ~ Streak)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Streak)), datax=TRUE, lwd=3, col="red")
```



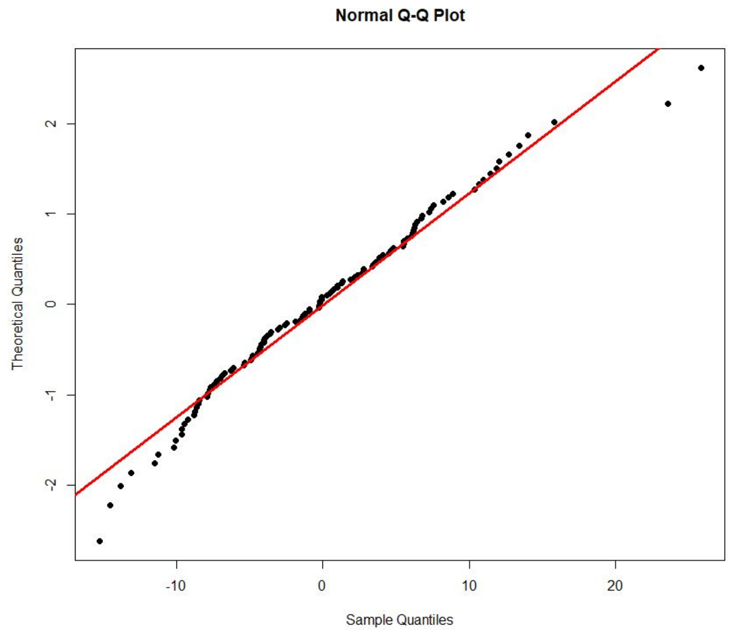
For Construction

```
> qqnorm(resid(lm(GmSc ~ Construction)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Construction)), datax=TRUE, lwd=3, col="red")
```



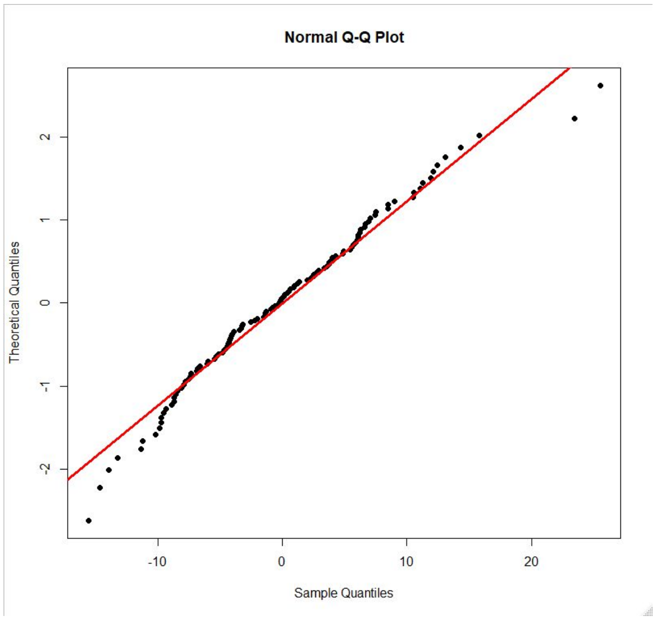
For Innovation:

```
> qqnorm(resid(lm(GmSc ~ Innovation)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Innovation)), datax=TRUE, lwd=3, col="red")
```



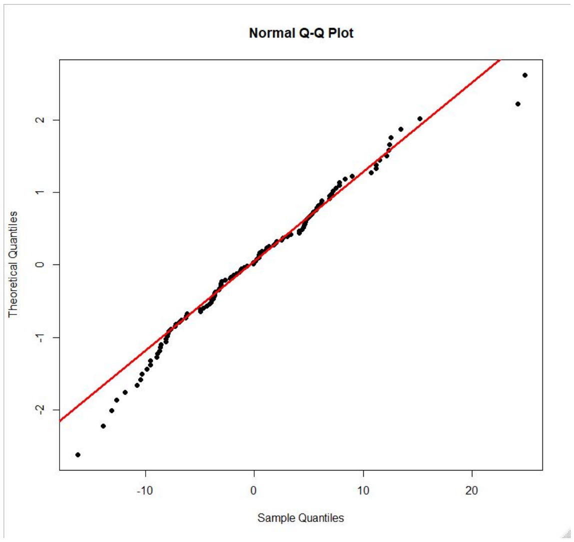
For Murder:

```
> qqnorm(resid(lm(GmSc ~ Murder)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Murder)), datax=TRUE, lwd=3, col="red")
```



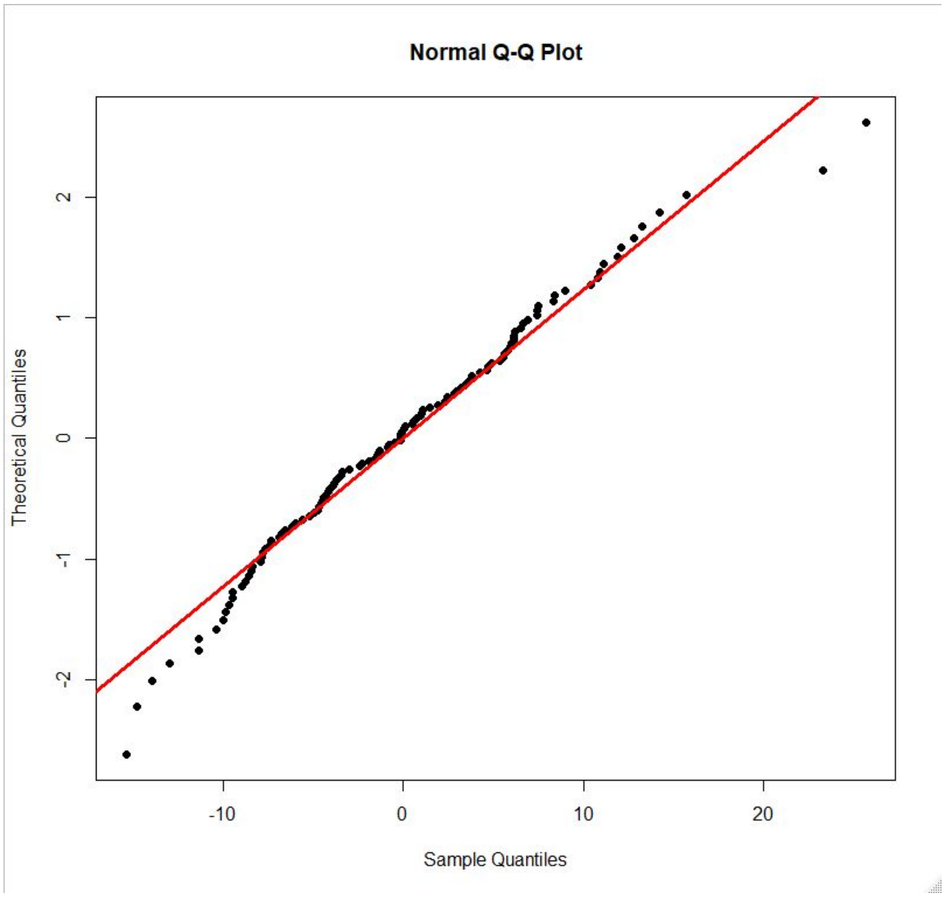
For Restaurants:

```
> qqnorm(resid(lm(GmSc ~ Restaurants)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Restaurants)), datax=TRUE, lwd=3, col="red")
```



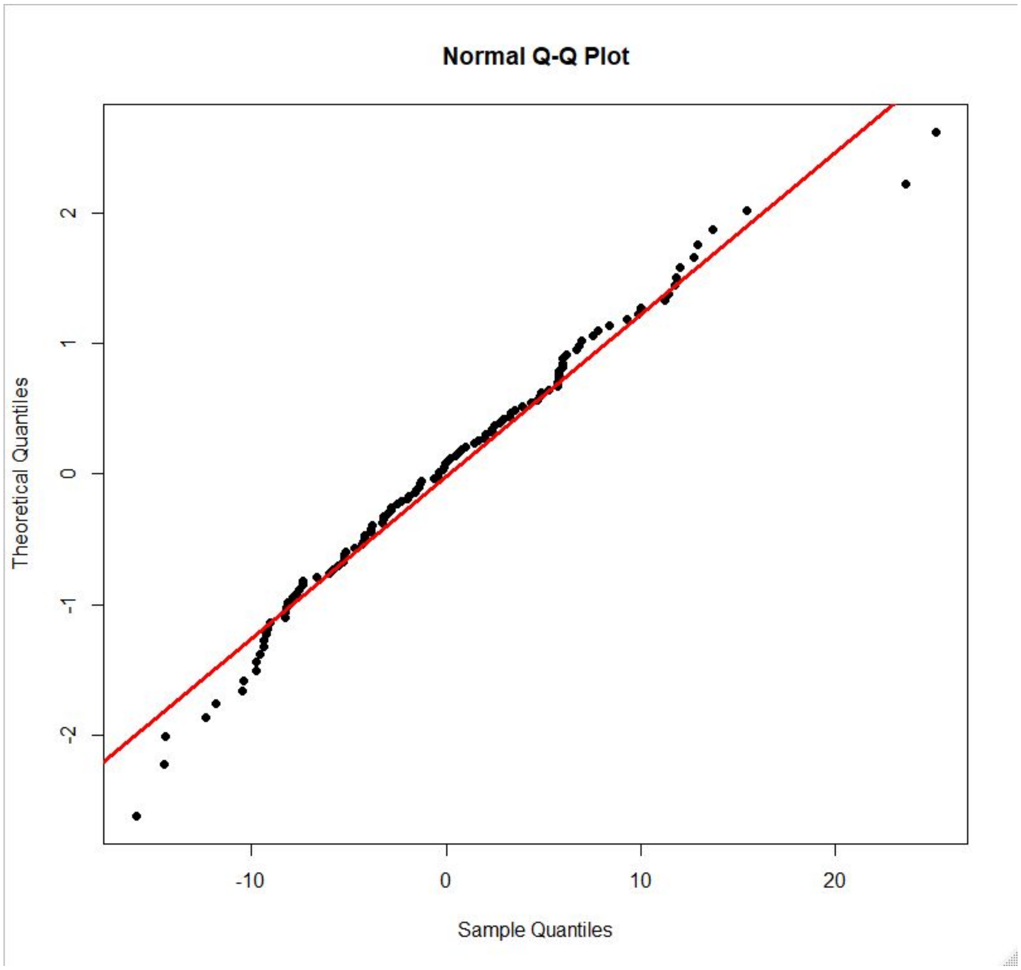
For Marijuana:

```
> qqnorm(resid(lm(GmSc ~ Marijuana)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Marijuana)), datax=TRUE, lwd=3, col="red")
```



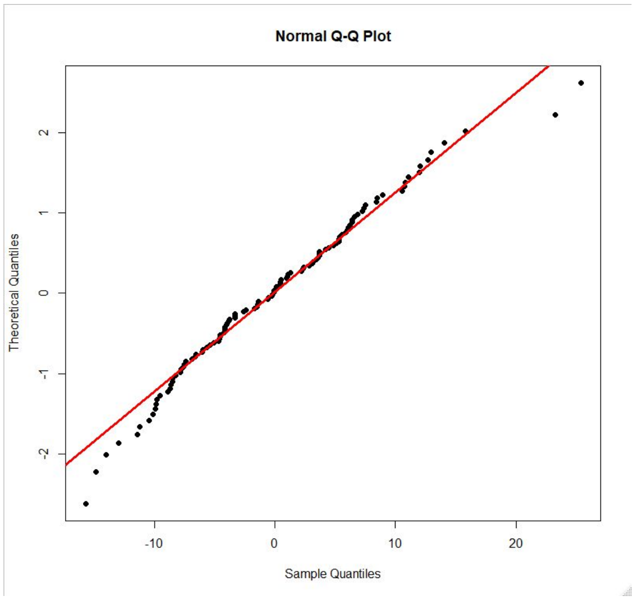
For Amusement Parks:

```
> qqnorm(resid(lm(GmSc ~ AmusementParks)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ AmusementParks)), datax=TRUE, lwd=3, col="red")
```



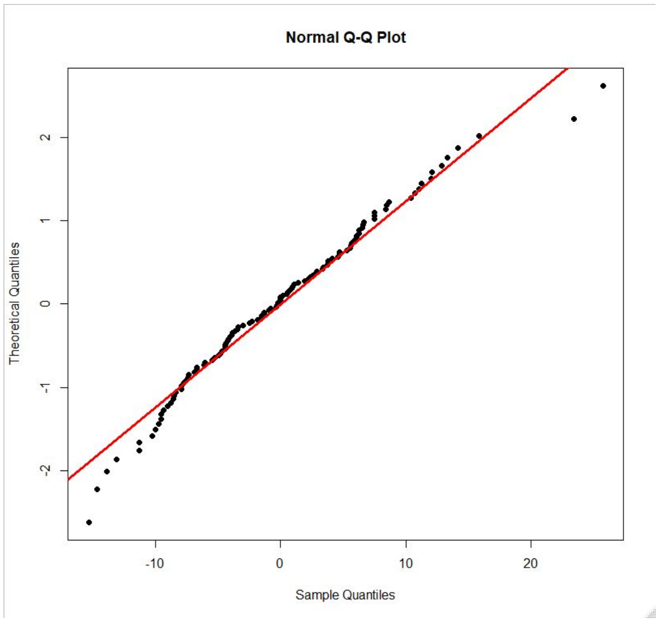
For Bars:

```
> qqnorm(resid(lm(GmSc ~ Bars)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Bars)), datax=TRUE, lwd=3, col="red")
```



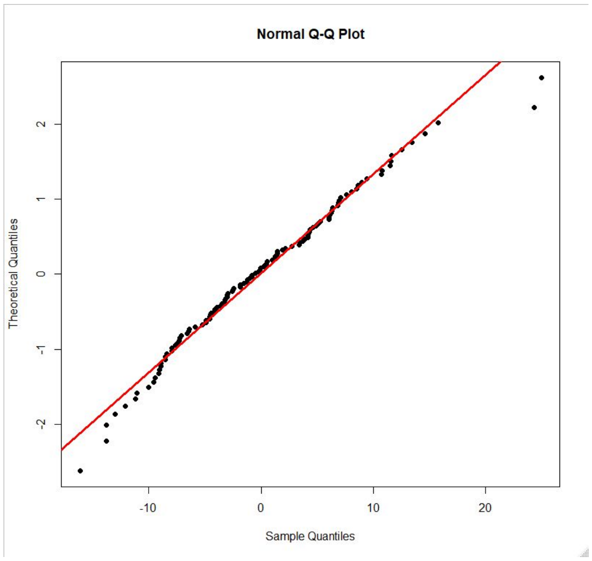
For Fun:

```
> qqnorm(resid(lm(GmSc ~ Fun)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Fun)), datax=TRUE, lwd=3, col="red")
```



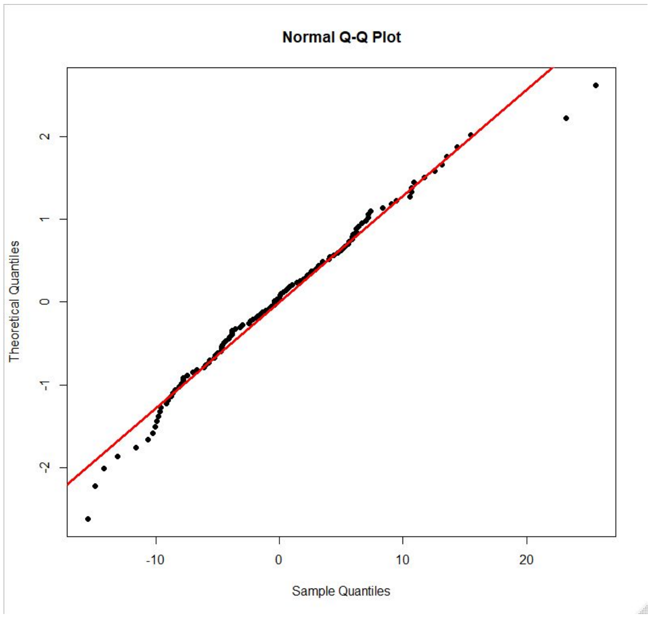
For Casinos:

```
> qqnorm(resid(lm(GmSc ~ Casinos)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Casinos)), datax=TRUE, lwd=3, col="red")
```



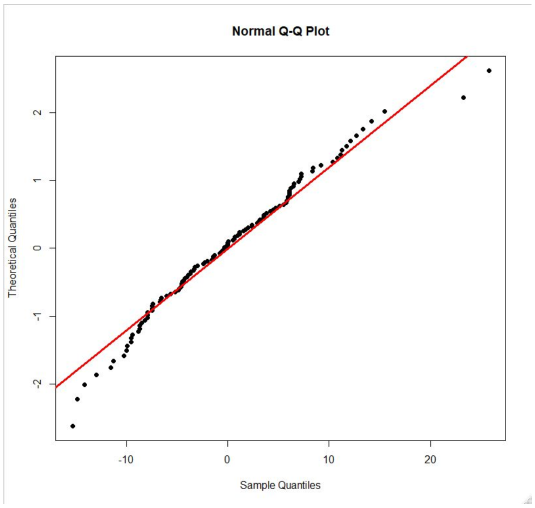
For Altitude:

```
> qqnorm(resid(lm(GmSc ~ Altitude)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ Altitude)), datax=TRUE, lwd=3, col="red")
```



For Acceleration of Gravity:

```
> qqnorm(resid(lm(GmSc ~ GravitationalConstant)), datax=TRUE, pch=16)
> qqline(resid(lm(GmSc ~ GravitationalConstant)), datax=TRUE, lwd=3, col="red")
```



**Anderson Darling Test for Normality in R with Code/Output Pairs for Indep. Variables: NOTE: These all pass the anderson darling test with alpha set to .01,.05, or even .1, meaning that the normality assumption holds or is reasonable.**

```
> ad.test(resid(lm(GmSc ~ AvgTemp)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ AvgTemp))
A = 0.41867, p-value = 0.3227

> ad.test(resid(lm(GmSc ~ Streak)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Streak))
A = 0.46641, p-value = 0.2476

> ad.test(resid(lm(GmSc ~ Construction)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Construction))
A = 0.40808, p-value = 0.3419

> ad.test(resid(lm(GmSc ~ Innovation)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Innovation))
A = 0.4394, p-value = 0.2879

> ad.test(resid(lm(GmSc ~ Murder)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Murder))
A = 0.43277, p-value = 0.2986
```

```
> ad.test(resid(lm(GmSc ~ Restaurants)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Restaurants))
A = 0.48546, p-value = 0.2224

> ad.test(resid(lm(GmSc ~ Marijuana)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Marijuana))
A = 0.41533, p-value = 0.3286

> ad.test(resid(lm(GmSc ~ AmusementParks)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ AmusementParks))
A = 0.41934, p-value = 0.3215

> ad.test(resid(lm(GmSc ~ Bars)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Bars))
A = 0.41415, p-value = 0.3308

> ad.test(resid(lm(GmSc ~ Fun)))

        Anderson-Darling normality test

data:  resid(lm(GmSc ~ Fun))
A = 0.43996, p-value = 0.287

> ad.test(resid(lm(GmSc ~ Casinos)))
```

```
	Anderson-Darling normality test

data:  resid(lm(GmSc ~ Casinos))
A = 0.40969, p-value = 0.3389

> ad.test(resid(lm(GmSc ~ Altitude)))

	Anderson-Darling normality test

data:  resid(lm(GmSc ~ Altitude))
A = 0.37442, p-value = 0.4101

> ad.test(resid(lm(GmSc ~ GravitationalConstant)))

	Anderson-Darling normality test

data:  resid(lm(GmSc ~ GravitationalConstant))
A = 0.42637, p-value = 0.3093
```

**R Libraries Used**

Leaps, Car, Nortest