Two correlation Plots of Happiness Alcohol Consumption:

Raghunath Reddy Koilakonda

4/6/2021

**Importing the Data and Looking at the Summary**

**Variable Description:**

Country - name of the country

Region - region the country belongs

Hemisphere - hemisphere of the country

HappinessScore - rate of happiness

HDI - Human Development Index

GDP\_PerCapita - Gross Domestic Product index (per capita)

Beer\_PerCapita - Liters of beer consumptions (per capita)

Wine\_PerCapita - Wine Consumption (per capita)

Spirit\_PerCapita - Consumption of spirits drink (per capita)

**For this Analysis, we will be using HappinessScore as our Dependent Variable to see if Alcohol Consumption influences HappinessScore.**

## Country Region Hemisphere HappinessScore   
## Length:122 Length:122 Length:122 Min. :3.069   
## Class :character Class :character Class :character 1st Qu.:4.528   
## Mode :character Mode :character Mode :character Median :5.542   
## Mean :5.525   
## 3rd Qu.:6.477   
## Max. :7.526   
## HDI GDP\_PerCapita Beer\_PerCapita Spirit\_PerCapita  
## Min. :351.0 Min. : 1.029 Min. : 1.00 Min. : 1.0   
## 1st Qu.:663.8 1st Qu.: 4.134 1st Qu.: 38.25 1st Qu.: 25.5   
## Median :757.5 Median : 12.016 Median :125.50 Median : 82.5   
## Mean :740.9 Mean : 91.483 Mean :137.57 Mean : 96.6   
## 3rd Qu.:861.5 3rd Qu.: 41.990 3rd Qu.:224.75 3rd Qu.:142.5   
## Max. :951.0 Max. :953.000 Max. :376.00 Max. :373.0   
## Wine\_PerCapita   
## Min. : 1.0   
## 1st Qu.: 5.0   
## Median : 16.0   
## Mean : 66.6   
## 3rd Qu.:112.8   
## Max. :370.0

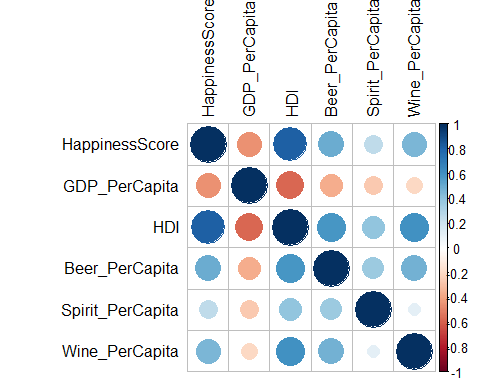
**Preparing the Data:**

#change the name to shorter  
hap = HappinessAlcoholConsumption  
#check to see if there is any na values  
colSums(is.na(hap))

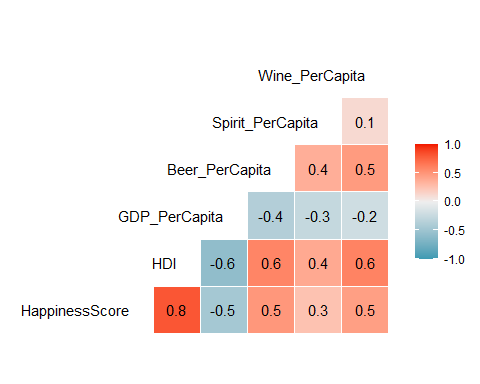
## Country Region Hemisphere HappinessScore   
## 0 0 0 0   
## HDI GDP\_PerCapita Beer\_PerCapita Spirit\_PerCapita   
## 0 0 0 0   
## Wine\_PerCapita   
## 0

**Correlogram for all observations**

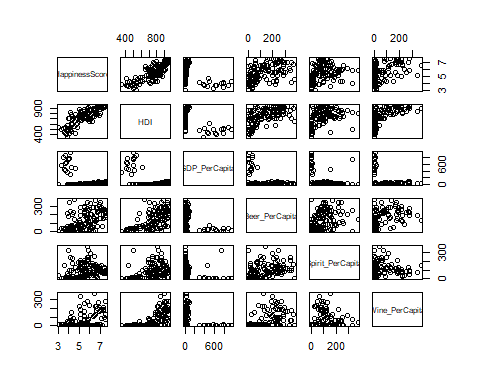
1. We used corrplot for the correlation matrix, and only used numeric values.



1. Values of the correlation:



Scatter Plot Matrix of the Variables to get a different look of the correlation.



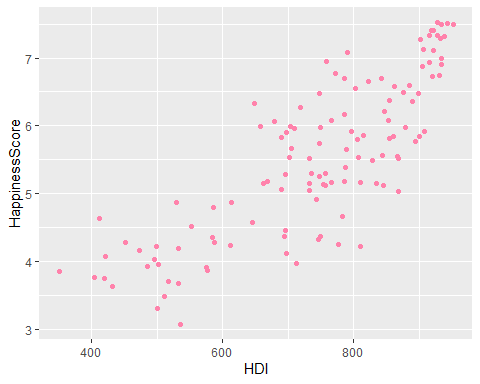
**Interpretations:**

* The independent variables that appear to be the most strongly positively correlated with the Happiness Score for each country : HDI(human development index), Beer\_PerCapita, Wine\_PerCapita
* The independent variables that appear to be the most strongly negatively correlated with the Happiness Score for each country: GDP\_PerCapita
* Beer\_PerCapita and Wine\_PerCapita have somewhat strong correlation to Happiness score and to each other as well as to HDI. So, we should take note of this in our heads to see if this could cause multicollinearity.

**Plots of the Data:**

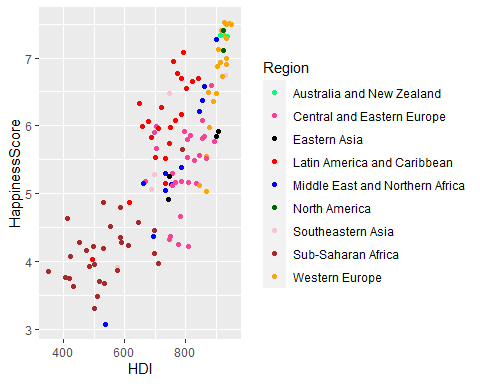
**HDI to Happiness Score:**

HDI and Happiness Score have a strong positive correlation. As HDI increases happiness score increases.

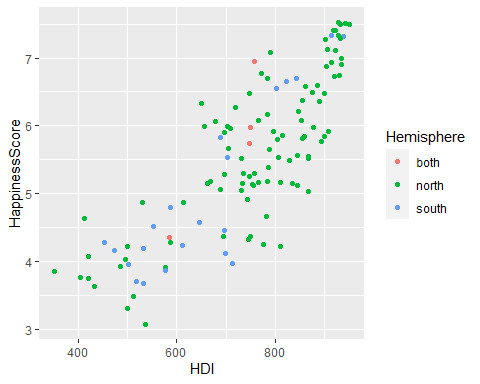


**HDI to Happiness Score with Region Involved:**

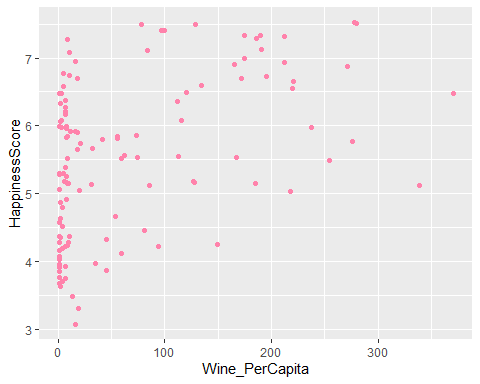
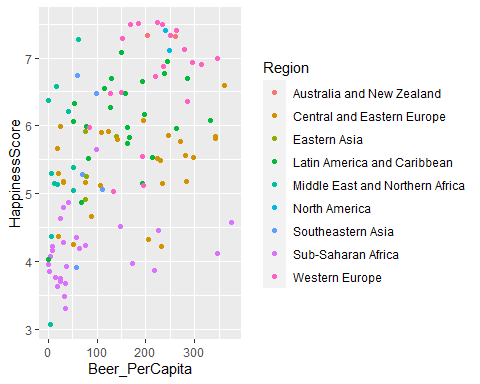
We seperate into colors based on Regions, you can tell most regions are in a specific clumps together. Where Western Europe has a high HDI and Happiness Score and Sub-Saharan Africa has a low HDI and low Happiness Score.



**HDI to Happiness Score with Hemisphere Emphasis:**

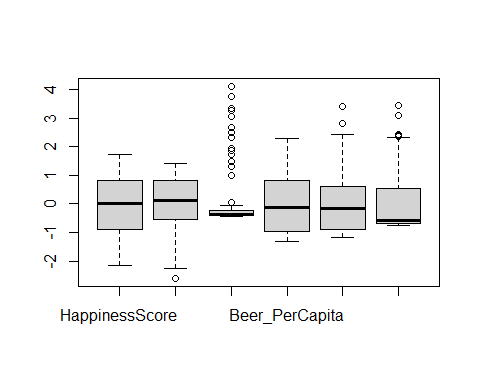
Based on the Scatter plot using the color for hemisphere there is no real correlation or difference on hemisphere

**Wine\_PerCapita to Happiness Score**

   
 **Beer\_PerCapita to Happiness Score Based on Region** 

**BoxPlot of all Numeric Values:**

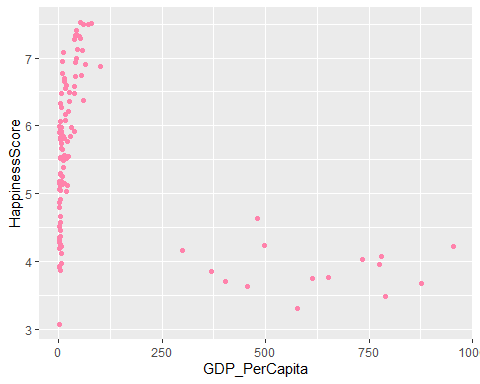
We can tell that there are lots of outliers for GDP\_PerCapita meaning most countries have a lower GDP than a higher one. Spirit\_percapita and wine\_percapita have outliers but not as many as GDP\_percapita, so we will look at a scatterplot with GDP per capita and see how many outliers



**Outliers of GDP\_PerCapita:**

As we can tell from the scatterplot, there are 15 points that are above 250 with outliers.

We can take out the 15 values that are outliers for the GDP per capita.



Take out 15 data points to see where they come from. So, based on this all of the outliers are in the sub-saharan africa region, if the models turn to be bad with error terms we will take out the 15 values.

**GDP\_PerCapia without 15 points**

## # A tibble: 15 x 9  
## Country Region Hemisphere HappinessScore HDI GDP\_PerCapita Beer\_PerCapita  
## <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Sierra~ Sub-S~ north 4.64 413 481 25  
## 2 Rep. C~ Sub-S~ south 4.24 612 498 76  
## 3 Senegal Sub-S~ north 4.22 499 953 9  
## 4 Malawi Sub-S~ south 4.16 474 300 8  
## 5 Mali Sub-S~ north 4.07 421 780 5  
## 6 Haiti Latin~ north 4.03 496 735 1  
## 7 Comoros Sub-S~ south 3.96 502 775 1  
## 8 Niger Sub-S~ north 3.86 351 368 3  
## 9 Chad Sub-S~ north 3.76 405 651 15  
## 10 Burkin~ Sub-S~ north 3.74 420 614 25  
## 11 Madaga~ Sub-S~ south 3.70 517 402 26  
## 12 Tanzan~ Sub-S~ south 3.67 533 878 36  
## 13 Liberia Sub-S~ north 3.62 432 455 19  
## 14 Benin Sub-S~ north 3.48 512 789 34  
## 15 Togo Sub-S~ north 3.30 500 577 36  
## # ... with 2 more variables: Spirit\_PerCapita <dbl>, Wine\_PerCapita <dbl>

**Linear Regression Model:**

Take out country variable and change region and hemisphere variables to factors and not characters.

**Model1 with all Values:**

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = hap)  
##   
## Coefficients:  
## (Intercept) RegionCentral and Eastern Europe   
## 1.863e+00 -1.476e+00   
## RegionEastern Asia RegionLatin America and Caribbean   
## -1.778e+00 -4.376e-01   
## RegionMiddle East and Northern Africa RegionNorth America   
## -1.495e+00 -4.303e-01   
## RegionSoutheastern Asia RegionSub-Saharan Africa   
## -1.173e+00 -1.173e+00   
## RegionWestern Europe Hemispherenorth   
## -6.920e-01 6.914e-02   
## Hemispheresouth HDI   
## -1.728e-01 6.665e-03   
## GDP\_PerCapita Beer\_PerCapita   
## 4.962e-05 -4.401e-04   
## Spirit\_PerCapita Wine\_PerCapita   
## -7.476e-04 -1.993e-03

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = hap)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.41669 -0.28443 0.05566 0.34364 1.11667   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.863e+00 8.342e-01 2.233 0.027640  
## RegionCentral and Eastern Europe -1.476e+00 4.503e-01 -3.277 0.001419  
## RegionEastern Asia -1.778e+00 5.250e-01 -3.387 0.000992  
## RegionLatin America and Caribbean -4.376e-01 4.494e-01 -0.974 0.332426  
## RegionMiddle East and Northern Africa -1.495e+00 4.824e-01 -3.098 0.002491  
## RegionNorth America -4.303e-01 5.800e-01 -0.742 0.459825  
## RegionSoutheastern Asia -1.173e+00 5.140e-01 -2.281 0.024543  
## RegionSub-Saharan Africa -1.173e+00 4.905e-01 -2.392 0.018526  
## RegionWestern Europe -6.920e-01 4.465e-01 -1.550 0.124154  
## Hemispherenorth 6.914e-02 2.760e-01 0.250 0.802686  
## Hemispheresouth -1.728e-01 2.948e-01 -0.586 0.558874  
## HDI 6.665e-03 8.191e-04 8.137 8.31e-13  
## GDP\_PerCapita 4.962e-05 3.245e-04 0.153 0.878775  
## Beer\_PerCapita -4.401e-04 7.331e-04 -0.600 0.549583  
## Spirit\_PerCapita -7.476e-04 8.710e-04 -0.858 0.392631  
## Wine\_PerCapita -1.993e-03 9.715e-04 -2.051 0.042701  
##   
## (Intercept) \*   
## RegionCentral and Eastern Europe \*\*   
## RegionEastern Asia \*\*\*  
## RegionLatin America and Caribbean   
## RegionMiddle East and Northern Africa \*\*   
## RegionNorth America   
## RegionSoutheastern Asia \*   
## RegionSub-Saharan Africa \*   
## RegionWestern Europe   
## Hemispherenorth   
## Hemispheresouth   
## HDI \*\*\*  
## GDP\_PerCapita   
## Beer\_PerCapita   
## Spirit\_PerCapita   
## Wine\_PerCapita \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5475 on 106 degrees of freedom  
## Multiple R-squared: 0.801, Adjusted R-squared: 0.7728   
## F-statistic: 28.44 on 15 and 106 DF, p-value: < 2.2e-16

Based on the full model(Model1) HDI, Wine PerCapita, and certain regions are statistically significant. As well as a high squared Showing the model is explained by most of the data.

**R-Squared Interpretation**

* 80% of the variability of Happiness Score can be explained by the model.

**P-value Interpretation** - After controlling for all the other independent variables in the model, there is a statistically significant difference between Region of Australia & New Zealand (intercept dummy variable) and Region: Central and Eastern Europe, Eastern Asia, East and Northern Africa, Southeastern Asia, Sub-Saharan Africa that effects the Happiness Score. The p-value is less than 0.05 on all these data points. So, these regions effect Happiness Score differently than Region of Australia and New Zealand.

* After controlling for Region and Wine\_PerCapita there is a statistically significant difference with HDI on Happiness Score.
* After controlling for Region and HDI there is a statistically significant difference with Wine\_PerCapita on Happiness Score.

**Coefficient Interpretations**

After controlling for hemisphere, HDI, GDP\_PerCapita, Beer\_PerCapita, Spirit\_PerCapita and Wine\_PerCapita, the Happiness Score for Region Central and Eastern Europe is about 1.476e less than the Happiness Score for Region Australia and New Zealand.

After controlling for region, HDI, GDP\_PerCapita, Beer\_PerCapita, Spirit\_PerCapita and Wine\_PerCapita, the Happiness Score for Hemisphere North is 6.9e-02 greater than Hemisphere Both.

After controlling for region, HDI, GDP\_PerCapita, Beer\_PerCapita, Spirit\_PerCapita and Wine\_PerCapita, the Happiness Score for Hemisphere South is 1.72e-1 lesser than Hemisphere Both.

**Error Assumption Plots**:

based on the error term plots we see the plots are normally distributed in the Normal Q-Q We see in the residuals vs Leverage there are some outliers but none in the 0.5 so this does not make it too significant

**Residual vs Fitted**: The points seem to be near the center line although the red line is sloping upward like a “U” a little. The red line almost follows dotted line 0. So, the error terms are linear.

**Normal Q-Q Plot:**

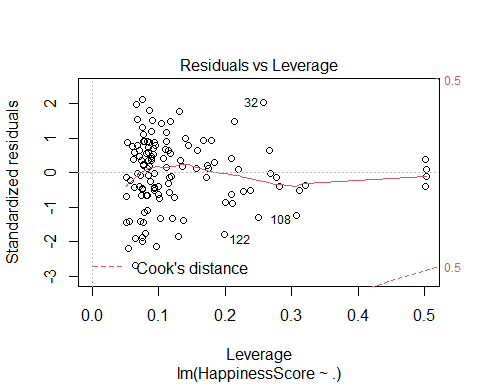
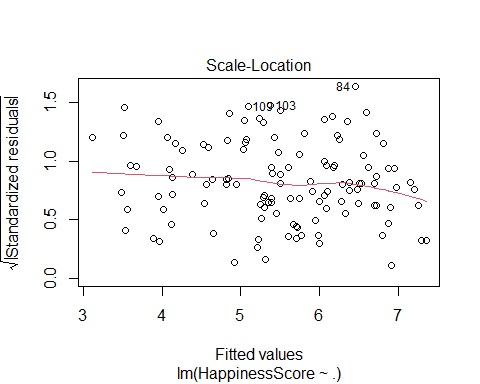
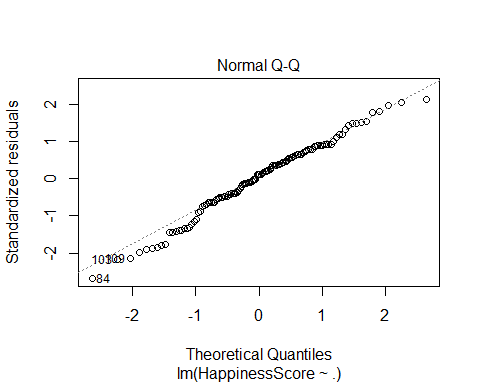
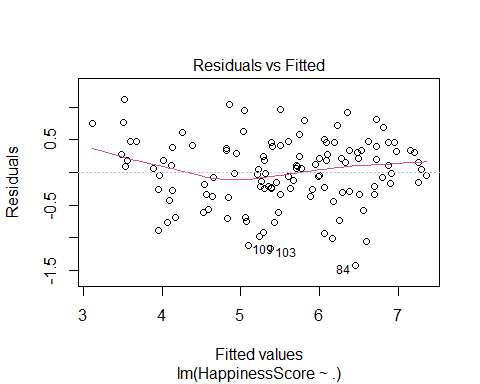
Dots are generally following the y = x line. So, the error terms follow a normal distribution.

**Scale Location Plot:**

We have constant variance because the red line is following y = 0 is in another “U” formation but not bad enough to where we can’t accept this model.

**Residuals vs. Leverage:**

Cooks distance doesn’t have any values in 0.5 or 1 so nothing is significantly influential.



**Multicollinearity: (model1)**

If we consider, the VIF we see Region has a VIF of 26 which means this variable is highly correlated to another variable most likely hemisphere. We will reduce the model and find the best model for model 1 using STEPWISE reduction.

## GVIF Df GVIF^(1/(2\*Df))  
## Region 26.844354 8 1.228297  
## Hemisphere 2.285345 2 1.229527  
## HDI 6.080761 1 2.465920  
## GDP\_PerCapita 1.867201 1 1.366455  
## Beer\_PerCapita 2.383472 1 1.543850  
## Spirit\_PerCapita 2.030639 1 1.425005  
## Wine\_PerCapita 2.958054 1 1.719900

**NewData Model- GDP <250**

We will create a model with the outliers of the 15 GDP variables to see if this makes a better model.

From the graph below you can see the 15 outliers that contribute to a not normal distribution.

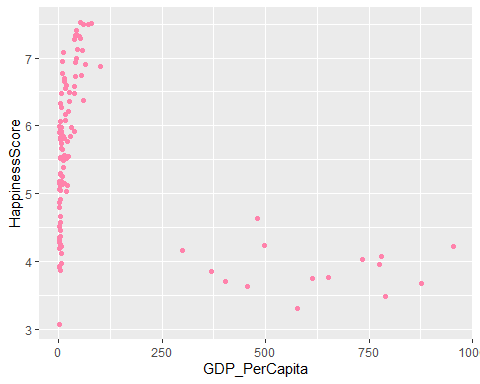
Comparing this model NewModel to Model1 we look at the rsquared and see Model1 has a better rsquared and more variables with low p-values. We also see the vif has gone down, but we are sacrificing a lower R-squared and less low p-values.

We decided based on the numbers we will stick with model1 with all 122 data points.

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = NewData)  
##   
## Coefficients:  
## (Intercept) Hemispherenorth Hemispheresouth HDI   
## 1.0063798 -0.4916810 -0.5671575 0.0068848   
## GDP\_PerCapita Beer\_PerCapita Spirit\_PerCapita Wine\_PerCapita   
## 0.0002859 0.0006511 -0.0014469 -0.0011076

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = NewData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5582 -0.5168 0.1020 0.5271 1.4074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0063798 0.5548903 1.814 0.0724 .   
## Hemispherenorth -0.4916810 0.3125866 -1.573 0.1185   
## Hemispheresouth -0.5671575 0.3402755 -1.667 0.0983 .   
## HDI 0.0068848 0.0007082 9.722 <2e-16 \*\*\*  
## GDP\_PerCapita 0.0002859 0.0003653 0.783 0.4354   
## Beer\_PerCapita 0.0006511 0.0007800 0.835 0.4056   
## Spirit\_PerCapita -0.0014469 0.0008746 -1.654 0.1008   
## Wine\_PerCapita -0.0011076 0.0009305 -1.190 0.2364   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6647 on 114 degrees of freedom  
## Multiple R-squared: 0.6846, Adjusted R-squared: 0.6652   
## F-statistic: 35.34 on 7 and 114 DF, p-value: < 2.2e-16

## GVIF Df GVIF^(1/(2\*Df))  
## Hemisphere 1.322906 2 1.072463  
## HDI 3.084156 1 1.756177  
## GDP\_PerCapita 1.604842 1 1.266824  
## Beer\_PerCapita 1.830707 1 1.353036  
## Spirit\_PerCapita 1.389436 1 1.178743  
## Wine\_PerCapita 1.841472 1 1.357008



**Stepwise for Model1:**

Using the stepwise function for forward and backward we get the equation with the lowest AIC: **HappinessScore ~ Region + HDI + Wine\_PerCapita**

## Start: AIC=-132.12  
## HappinessScore ~ Region + Hemisphere + HDI + GDP\_PerCapita +   
## Beer\_PerCapita + Spirit\_PerCapita + Wine\_PerCapita  
##   
## Df Sum of Sq RSS AIC  
## - Hemisphere 2 0.5100 32.287 -134.180  
## - GDP\_PerCapita 1 0.0070 31.784 -134.095  
## - Beer\_PerCapita 1 0.1080 31.885 -133.708  
## - Spirit\_PerCapita 1 0.2209 31.998 -133.277  
## <none> 31.777 -132.122  
## - Wine\_PerCapita 1 1.2615 33.039 -129.373  
## - Region 8 18.5888 50.366 -91.933  
## - HDI 1 19.8492 51.627 -74.918  
##   
## Step: AIC=-134.18  
## HappinessScore ~ Region + HDI + GDP\_PerCapita + Beer\_PerCapita +   
## Spirit\_PerCapita + Wine\_PerCapita  
##   
## Df Sum of Sq RSS AIC  
## - GDP\_PerCapita 1 0.0462 32.334 -136.005  
## - Beer\_PerCapita 1 0.1388 32.426 -135.657  
## - Spirit\_PerCapita 1 0.1961 32.484 -135.441  
## <none> 32.287 -134.180  
## - Wine\_PerCapita 1 1.7243 34.012 -129.833  
## - Region 8 19.3446 51.632 -92.905  
## - HDI 1 19.4814 51.769 -78.583  
##   
## Step: AIC=-136.01  
## HappinessScore ~ Region + HDI + Beer\_PerCapita + Spirit\_PerCapita +   
## Wine\_PerCapita  
##   
## Df Sum of Sq RSS AIC  
## - Beer\_PerCapita 1 0.1720 32.506 -137.358  
## - Spirit\_PerCapita 1 0.1759 32.510 -137.343  
## <none> 32.334 -136.005  
## - Wine\_PerCapita 1 1.6899 34.024 -131.790  
## - Region 8 19.4995 51.833 -94.431  
## - HDI 1 20.4132 52.747 -78.299  
##   
## Step: AIC=-137.36  
## HappinessScore ~ Region + HDI + Spirit\_PerCapita + Wine\_PerCapita  
##   
## Df Sum of Sq RSS AIC  
## - Spirit\_PerCapita 1 0.2683 32.774 -138.355  
## <none> 32.506 -137.358  
## - Wine\_PerCapita 1 1.7850 34.291 -132.836  
## - Region 8 19.6514 52.157 -95.671  
## - HDI 1 25.4997 58.005 -68.705  
##   
## Step: AIC=-138.36  
## HappinessScore ~ Region + HDI + Wine\_PerCapita  
##   
## Df Sum of Sq RSS AIC  
## <none> 32.774 -138.355  
## - Wine\_PerCapita 1 1.5977 34.372 -134.548  
## - Region 8 20.3841 53.158 -95.352  
## - HDI 1 25.2579 58.032 -70.650

**R-Squared Interpretation**

* 79% of the variability of Happiness Score can be explained by the model.

**P-value Interpretation** - After controlling for all the other independent variables in the model, there is a statistically significant difference between Region of Australia & New Zealand(intercept dummy variable) and Region: Central and Eastern Europe, Eastern Asia, East and Northern Africa, Southeastern Asia, Sub-Saharan Africa that effects the Happiness Score. The p-value is less than 0.05.

* After controlling for Region and Wine\_PerCapita there is a statistically significant difference with HDI on Happiness Score.
* After controlling for Region and HDI there is a statistically significant differnce with Wine\_PerCapita on Happiness Score.

**Coefficient Interpretations:**

* As we control for Region, and Wine\_PerCapita. If you increase HDI by 1 the Happiness Score goes up by 0.006.
* As we control for Region, HDI. If you increase Wine\_PerCapita by 1 the Happiness Score goes down by .002.
* As we control for HDI and Wine\_PerCapita and holding all other regions constant. An increase by 1 of Region Sub-Saharan will decrease the Happiness Score by 1.17, which is different than the intercept dummy of Region Australia.

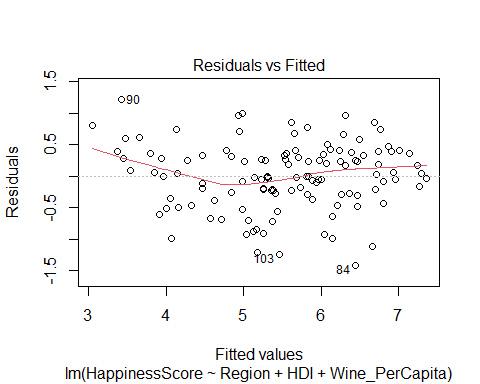
##   
## Call:  
## lm(formula = HappinessScore ~ Region + HDI + Wine\_PerCapita,   
## data = hap)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.40907 -0.28131 -0.00182 0.35635 1.21487   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.0812308 0.7015605 2.967 0.00369  
## RegionCentral and Eastern Europe -1.3694732 0.4139323 -3.308 0.00127  
## RegionEastern Asia -1.6145935 0.4978449 -3.243 0.00156  
## RegionLatin America and Caribbean -0.3877228 0.4334833 -0.894 0.37302  
## RegionMiddle East and Northern Africa -1.2561773 0.4520599 -2.779 0.00641  
## RegionNorth America -0.2584032 0.5511501 -0.469 0.64010  
## RegionSoutheastern Asia -1.0248544 0.4914656 -2.085 0.03933  
## RegionSub-Saharan Africa -1.1775230 0.4776040 -2.465 0.01521  
## RegionWestern Europe -0.4721578 0.4032105 -1.171 0.24411  
## HDI 0.0061034 0.0006599 9.249 1.92e-15  
## Wine\_PerCapita -0.0021317 0.0009164 -2.326 0.02182  
##   
## (Intercept) \*\*   
## RegionCentral and Eastern Europe \*\*   
## RegionEastern Asia \*\*   
## RegionLatin America and Caribbean   
## RegionMiddle East and Northern Africa \*\*   
## RegionNorth America   
## RegionSoutheastern Asia \*   
## RegionSub-Saharan Africa \*   
## RegionWestern Europe   
## HDI \*\*\*  
## Wine\_PerCapita \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5434 on 111 degrees of freedom  
## Multiple R-squared: 0.7947, Adjusted R-squared: 0.7762   
## F-statistic: 42.98 on 10 and 111 DF, p-value: < 2.2e-16

**Multicollinearity of the Best Model:**

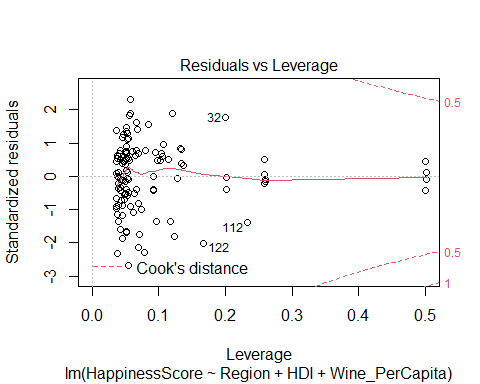
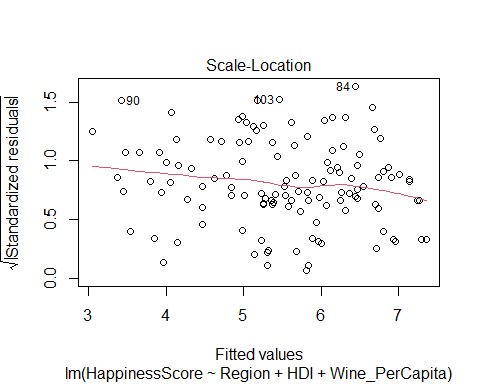
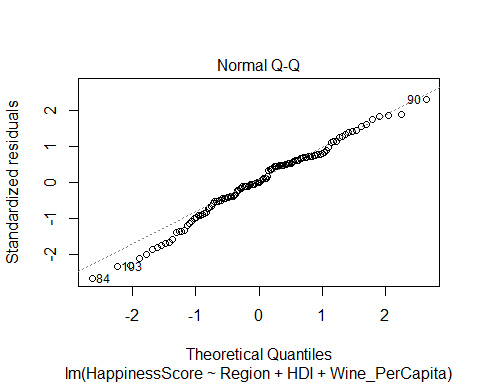
The VIF is lower than 10 so this means that there is no multicollinearity in the model.

## GVIF Df GVIF^(1/(2\*Df))  
## Region 6.286222 8 1.121758  
## HDI 4.006967 1 2.001741  
## Wine\_PerCapita 2.672467 1 1.634768

**All Subsets Model of FullModel(Model1)**

**Residual Plots of Best Model**

Based on the error term plots the data points are linear and normally distributed, with a constant variance.

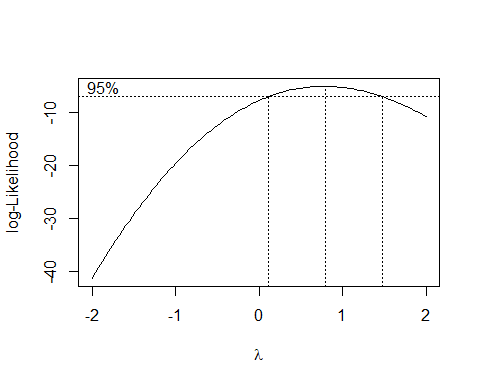


**BoxCox**

We are looking to see if the data needs to be transformed. Since 1 is in the CI we do not transform and 0 is not in so we keep the normal linear regression and do not do any transformation.

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select



**Train and Test Data**

Split data into train and test set approx. 80% in the train and test is 20%.

**Train Linear Regression Model**

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.32369 -0.33106 0.02631 0.39156 1.07743   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.935e+00 1.019e+00 1.899 0.0611  
## RegionCentral and Eastern Europe -1.422e+00 6.287e-01 -2.262 0.0264  
## RegionEastern Asia -1.555e+00 8.559e-01 -1.817 0.0729  
## RegionLatin America and Caribbean -2.697e-01 6.274e-01 -0.430 0.6684  
## RegionMiddle East and Northern Africa -1.507e+00 6.659e-01 -2.264 0.0262  
## RegionNorth America -3.265e-01 7.293e-01 -0.448 0.6556  
## RegionSoutheastern Asia -9.984e-01 6.822e-01 -1.463 0.1472  
## RegionSub-Saharan Africa -1.080e+00 6.688e-01 -1.614 0.1103  
## RegionWestern Europe -7.852e-01 6.227e-01 -1.261 0.2109  
## Hemispherenorth 1.006e-01 2.923e-01 0.344 0.7318  
## Hemispheresouth -2.381e-01 3.262e-01 -0.730 0.4675  
## HDI 6.325e-03 9.625e-04 6.571 4.3e-09  
## GDP\_PerCapita 2.767e-06 3.486e-04 0.008 0.9937  
## Beer\_PerCapita 3.308e-05 8.325e-04 0.040 0.9684  
## Spirit\_PerCapita -1.410e-03 1.095e-03 -1.288 0.2013  
## Wine\_PerCapita -1.054e-03 1.183e-03 -0.891 0.3757  
##   
## (Intercept) .   
## RegionCentral and Eastern Europe \*   
## RegionEastern Asia .   
## RegionLatin America and Caribbean   
## RegionMiddle East and Northern Africa \*   
## RegionNorth America   
## RegionSoutheastern Asia   
## RegionSub-Saharan Africa   
## RegionWestern Europe   
## Hemispherenorth   
## Hemispheresouth   
## HDI \*\*\*  
## GDP\_PerCapita   
## Beer\_PerCapita   
## Spirit\_PerCapita   
## Wine\_PerCapita   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.566 on 82 degrees of freedom  
## Multiple R-squared: 0.7988, Adjusted R-squared: 0.762   
## F-statistic: 21.71 on 15 and 82 DF, p-value: < 2.2e-16

**Multicollinearity:**

## GVIF Df GVIF^(1/(2\*Df))  
## Region 32.742976 8 1.243641  
## Hemisphere 2.409789 2 1.245933  
## HDI 6.406569 1 2.531120  
## GDP\_PerCapita 1.847878 1 1.359367  
## Beer\_PerCapita 2.474113 1 1.572931  
## Spirit\_PerCapita 2.339972 1 1.529697  
## Wine\_PerCapita 3.602980 1 1.898152

Based on the VIF looks like we could take ut the region since the vif = 32, We took out the region but the model adjusted r-squared went down to .65. So, we checked by taking out the hemisphere and keeping region since they are highly correlated. When we did that the adjusted r-squared went up to .76

**TrainModel2 without Hemisphere**

Taking out Hemisphere and rerunning a new Linear Regression Model

##   
## Call:  
## lm(formula = HappinessScore ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.33435 -0.35187 0.04453 0.35355 1.18376   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.068e+00 9.733e-01 2.125 0.0365  
## RegionCentral and Eastern Europe -1.186e+00 6.111e-01 -1.942 0.0555  
## RegionEastern Asia -1.378e+00 8.497e-01 -1.622 0.1086  
## RegionLatin America and Caribbean -1.551e-01 6.226e-01 -0.249 0.8039  
## RegionMiddle East and Northern Africa -1.330e+00 6.569e-01 -2.024 0.0461  
## RegionNorth America -5.287e-02 7.088e-01 -0.075 0.9407  
## RegionSoutheastern Asia -8.259e-01 6.741e-01 -1.225 0.2239  
## RegionSub-Saharan Africa -1.139e+00 6.692e-01 -1.702 0.0925  
## RegionWestern Europe -4.573e-01 5.865e-01 -0.780 0.4378  
## HDI 6.044e-03 9.459e-04 6.390 8.82e-09  
## GDP\_PerCapita 6.956e-05 3.449e-04 0.202 0.8406  
## Beer\_PerCapita -1.082e-05 8.337e-04 -0.013 0.9897  
## Spirit\_PerCapita -1.380e-03 1.097e-03 -1.259 0.2117  
## Wine\_PerCapita -1.536e-03 1.133e-03 -1.355 0.1791  
##   
## (Intercept) \*   
## RegionCentral and Eastern Europe .   
## RegionEastern Asia   
## RegionLatin America and Caribbean   
## RegionMiddle East and Northern Africa \*   
## RegionNorth America   
## RegionSoutheastern Asia   
## RegionSub-Saharan Africa .   
## RegionWestern Europe   
## HDI \*\*\*  
## GDP\_PerCapita   
## Beer\_PerCapita   
## Spirit\_PerCapita   
## Wine\_PerCapita   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5673 on 84 degrees of freedom  
## Multiple R-squared: 0.793, Adjusted R-squared: 0.7609   
## F-statistic: 24.75 on 13 and 84 DF, p-value: < 2.2e-16

**Finding BestModel for Train using the stepwise both approach**

##   
## Call:  
## lm(formula = HappinessScore ~ Region + HDI, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.36454 -0.32071 0.01858 0.36444 1.19792   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.9870534 0.8832207 2.250 0.0270  
## RegionCentral and Eastern Europe -1.1895945 0.5865311 -2.028 0.0456  
## RegionEastern Asia -1.2987936 0.8098951 -1.604 0.1124  
## RegionLatin America and Caribbean -0.0167549 0.5977955 -0.028 0.9777  
## RegionMiddle East and Northern Africa -1.0352398 0.6126261 -1.690 0.0946  
## RegionNorth America 0.0318477 0.6907270 0.046 0.9633  
## RegionSoutheastern Asia -0.6274883 0.6362242 -0.986 0.3267  
## RegionSub-Saharan Africa -0.8848651 0.6427229 -1.377 0.1721  
## RegionWestern Europe -0.5180055 0.5807734 -0.892 0.3749  
## HDI 0.0056780 0.0007247 7.835 1.01e-11  
##   
## (Intercept) \*   
## RegionCentral and Eastern Europe \*   
## RegionEastern Asia   
## RegionLatin America and Caribbean   
## RegionMiddle East and Northern Africa .   
## RegionNorth America   
## RegionSoutheastern Asia   
## RegionSub-Saharan Africa   
## RegionWestern Europe   
## HDI \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5639 on 88 degrees of freedom  
## Multiple R-squared: 0.7857, Adjusted R-squared: 0.7638   
## F-statistic: 35.84 on 9 and 88 DF, p-value: < 2.2e-16

**Model Evaluation**

Looking at the prediction for the step model “trainingmodel3” and the original “trainingmodel2”

Prediction = prediction with the all variables minus hemisphere Prediction2 = stepwise model

The first 1-24 is trainingmodel2, and the second 1-24 is trainingmodel3.

## 1 2 3 4 5 6 7 8   
## 7.047467 7.065555 6.826726 7.221103 6.082791 5.756667 6.001126 5.918196   
## 9 10 11 12 13 14 15 16   
## 5.870423 4.753524 6.094127 6.006785 5.166341 5.789223 5.587204 5.878664   
## 17 18 19 20 21 22 23 24   
## 5.385695 4.933364 4.781598 5.423391 6.124825 3.798000 4.266559 3.489596

## 1 2 3 4 5 6 7 8   
## 6.766605 6.868808 6.681435 7.182406 6.073353 5.846234 6.052767 5.825648   
## 9 10 11 12 13 14 15 16   
## 5.838189 4.755012 5.798443 5.882428 5.368234 5.873574 5.589675 5.956241   
## 17 18 19 20 21 22 23 24   
## 5.095691 4.935390 4.596029 5.396624 6.132259 3.793551 4.378384 3.486940

**Model Calculations**

Calculations of MAPE, RMSE, and MAE based on above model Evaluation

We may not have enough number of variables but we are checking to see if we can predict the model with only 122 observations

These two models MAE,RMSE, MAPE are very similar the step model is a bit better for these error values. Since, **the MAPE values are under 10% so this means the model is better able to forecast values.**

## MAE RMSE MAPE  
## all\_variables 0.4048331 0.5193684 0.06983226  
## withbetter\_model 0.4320188 0.5474118 0.07419036

**Heteroskedasticity**

The test statistic is 1.3 and the corresponding p-value is 0.86. Since the p-value is not less than 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say heteroskedasticity is present in the model.

##   
## studentized Breusch-Pagan test  
##   
## data: trainingmodel3  
## BP = 10.3, df = 9, p-value = 0.3268

**Normality**

The p-value is greater than 0.05 so we fail to reject the Shapiro test(normality). We can say the model is normally distributed based on the p-value is 0.4.

##   
## Shapiro-Wilk normality test  
##   
## data: trainingmodel3$residuals  
## W = 0.98555, p-value = 0.3616

**Multicollinearity**

Vif in both models is below 10 so we can conclude that multicollinearity of models is small. So, these independent variables in the two models are not very similar.

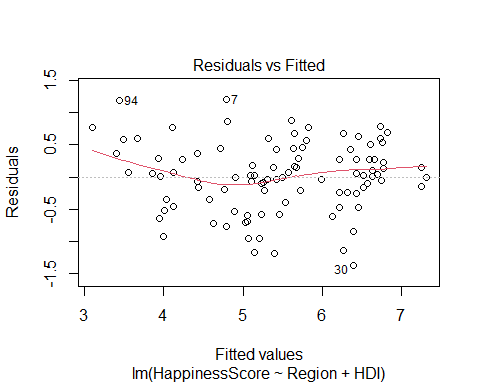
## GVIF Df GVIF^(1/(2\*Df))  
## Region 17.809285 8 1.197196  
## HDI 6.159112 1 2.481756  
## GDP\_PerCapita 1.800180 1 1.341708  
## Beer\_PerCapita 2.470161 1 1.571675  
## Spirit\_PerCapita 2.339158 1 1.529430  
## Wine\_PerCapita 3.288751 1 1.813491

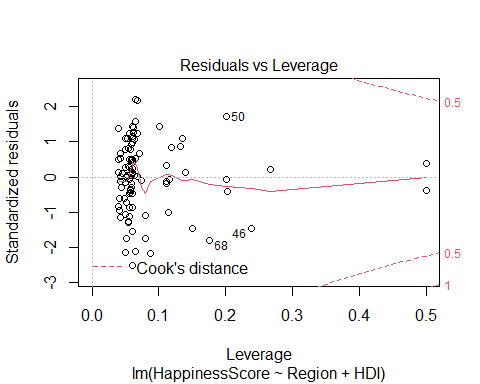
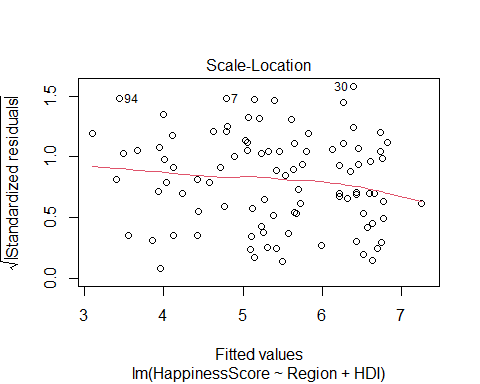
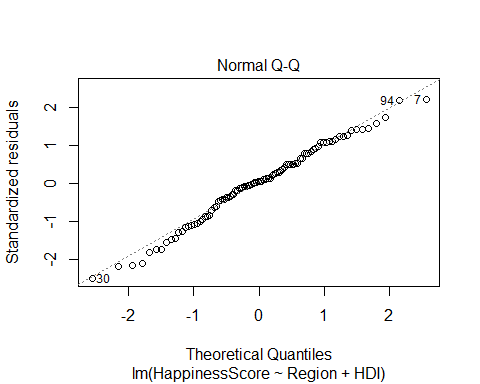
## GVIF Df GVIF^(1/(2\*Df))  
## Region 3.658971 8 1.084451  
## HDI 3.658971 1 1.912844

## Warning: not plotting observations with leverage one:  
## 13, 79

**Error Assumption Plots**:

Based on the error term plots the data points are linear and normally distributed, with a constant variance.





**Conclusion:**

Happiness Score is more highly correlated with other factors than alcohol consumption. HDI was the most highly correlated and related to Happiness Score. It would be interesting to see if weather would affect happiness and how many times it rains as well as the activity rate in activity in each country. We also would think looking at different cities in each country would be interesting to dive deeper in.