DSC 465 Project Deliverable 2

Name: Vismay Patel

Summary stats:

```
> summary(HP)
      Year             Country                                  Region
 Min.   :2015   Afghanistan:  5   Sub-Saharan Africa              :195
 1st Qu.:2016   Albania    :  5   Central and Eastern Europe      :145
 Median :2017   Algeria    :  5   Latin America and Caribbean     :111
 Mean   :2017   Argentina  :  5   Western Europe                  :105
 3rd Qu.:2018   Armenia    :  5   Middle East and Northern Africa: 97
 Max.   :2019   Australia  :  5   Southeastern Asia               : 54
                (Other)    :752   (Other)                        : 75
  Overall_Rank   Happiness_Score Social_Support    Economy_GDP         Health
 Min.   :  1.0   Min.   :2.693   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 40.0   1st Qu.:4.510   1st Qu.:0.8694   1st Qu.:0.6065   1st Qu.:0.4402
 Median : 79.0   Median :5.322   Median :1.1247   Median :0.9822   Median :0.6473
 Mean   : 78.7   Mean   :5.379   Mean   :1.0784   Mean   :0.9160   Mean   :0.6124
 3rd Qu.:118.0   3rd Qu.:6.189   3rd Qu.:1.3273   3rd Qu.:1.2362   3rd Qu.:0.8080
 Max.   :158.0   Max.   :7.769   Max.   :1.6440   Max.   :2.0960   Max.   :1.1410

    Freedom         Generosity     Trust_Corruption Dystopia_Residual
 Min.   :0.0000   Min.   :0.0000   Min.   :  1.0    Min.   :0.3286
 1st Qu.:0.3098   1st Qu.:0.1300   1st Qu.:143.2    1st Qu.:1.7380
 Median :0.4310   Median :0.2020   Median :287.5    Median :2.0946
 Mean   :0.4111   Mean   :0.2186   Mean   :299.6    Mean   :2.0927
 3rd Qu.:0.5310   3rd Qu.:0.2788   3rd Qu.:452.8    3rd Qu.:2.4556
 Max.   :0.7240   Max.   :0.8381   Max.   :636.0    Max.   :3.8377
                                                    NA's   :312
```

Figure 1.0

Here, in the summary stats we can see that the mean and median of all the numeric variables are similar. This shows that the data is symmetric.

Structure Analysis:

```
> HP$Trust_Corruption = as.numeric(HP$Trust_Corruption)
> str(HP)                  # Verifying if the change has taken place.
'data.frame':   782 obs. of  12 variables:
 $ Year             : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
 $ Country          : Factor w/ 166 levels "Afghanistan",..: 144 61 39 110 26 47 103 143 104 7 ...
 $ Region           : Factor w/ 10 levels "Australia and New Zealand",..: 10 10 10 10 6 10 10 10 1 1 ...
 $ Overall_Rank     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Happiness_Score  : num  7.59 7.56 7.53 7.52 7.43 ...
 $ Social_Support   : num  1.35 1.4 1.36 1.33 1.32 ...
 $ Economy_GDP      : num  1.4 1.3 1.33 1.46 1.33 ...
 $ Health           : num  0.941 0.948 0.875 0.885 0.906 ...
 $ Freedom          : num  0.666 0.629 0.649 0.67 0.633 ...
 $ Generosity       : num  0.297 0.436 0.341 0.347 0.458 ...
 $ Trust_Corruption : num  619 420 631 594 585 617 577 621 620 591 ...
 $ Dystopia_Residual: num  2.52 2.7 2.49 2.47 2.45 ...
```

Figure 1.1

The structure of the dataset initially consisted of three factors, Country, Region and Trust_corruption. But in since trust_corruption consists of numeric values it was changed to numeric type.

Now, we have two factors, one integer value and eight numeric values.

Exploratory visualizations of the hypothesis:

Hypothesis 1:

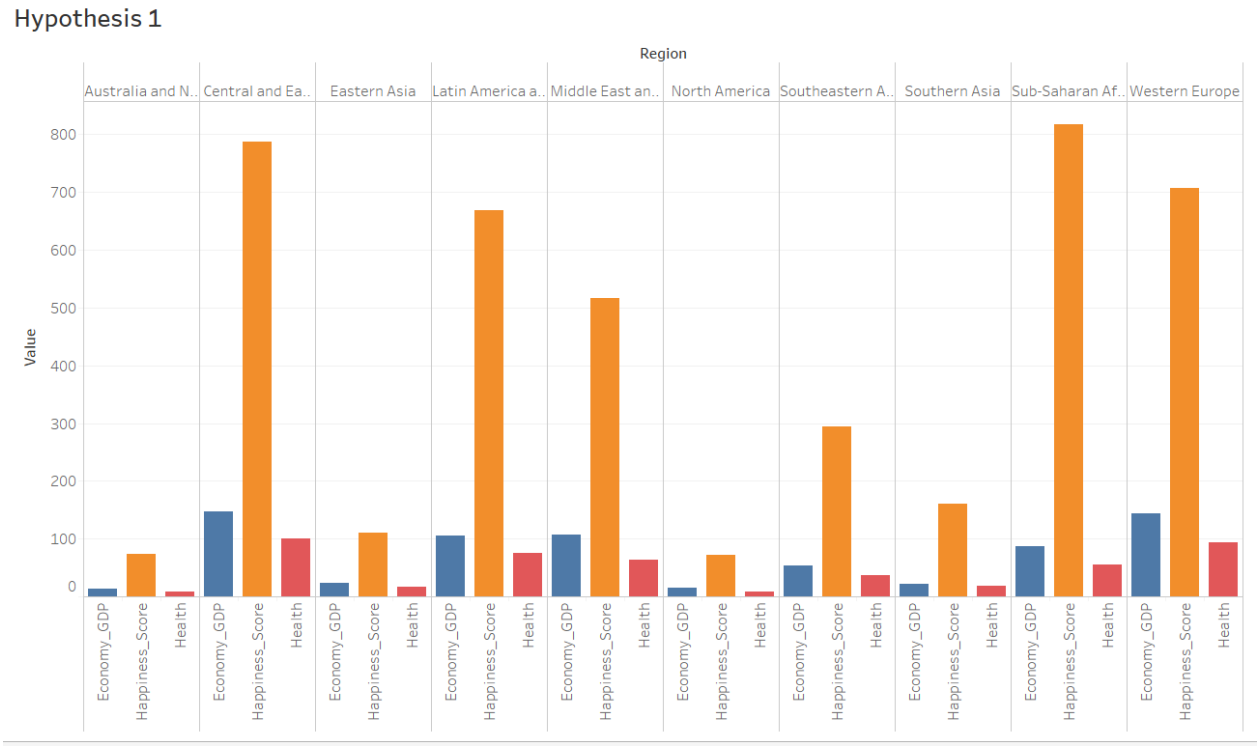

Figure 1.2

The first hypothesis says “Happiness score is higher in regions/countries having good Economic GDP and Health”. From this plot we can visualize that the economically strong and healthy countries have higher happiness scores.

Hypothesis 2:

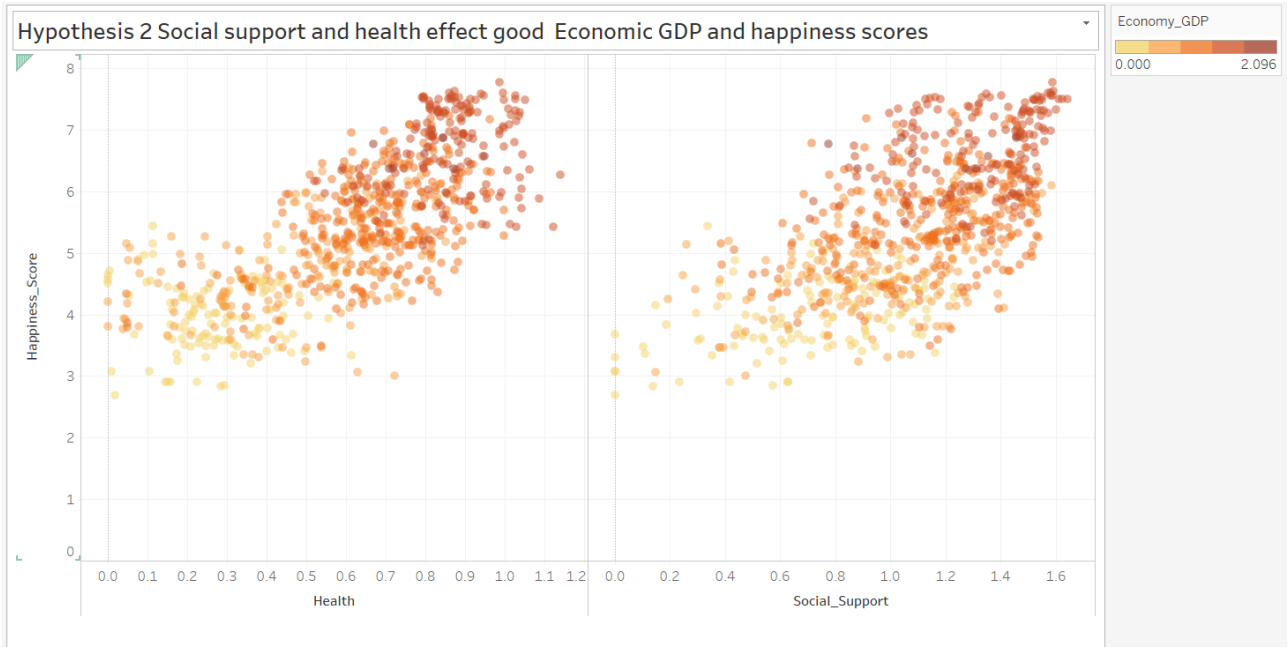

Figure 1.3

In the second hypothesis I am visualizing that social support and health effect the Economic GDP and Happiness scores. From the visualization it is clear that, economically strong countries have a higher happiness score. This is shown in with a scatterplot and the filter on the right, with happiness scores on y axis.