Name: Ishita More
UID: 2019130039
TE Comps
Subject: Data Analytics

ISE-Part3

Aim:

Correlation and Regression

Introduction:

Probably the most common statistical procedures used are correlation and linear regression (for data measured on a continuous scale), and analysis of variance (ANOVA) for comparing mean responses among more than two treatment groups (for two treatment groups, use the familiar t-test or its nonparametric equivalent). Regression and ANOVA are usually discussed together because Fisher demonstrated that all degrees of freedom and sums of squares (i.e., deviations from overall or within-group means) in an ANOVA problem is reducible to single-degree-of-freedom contrasts analyzable by regression.

DataSet:

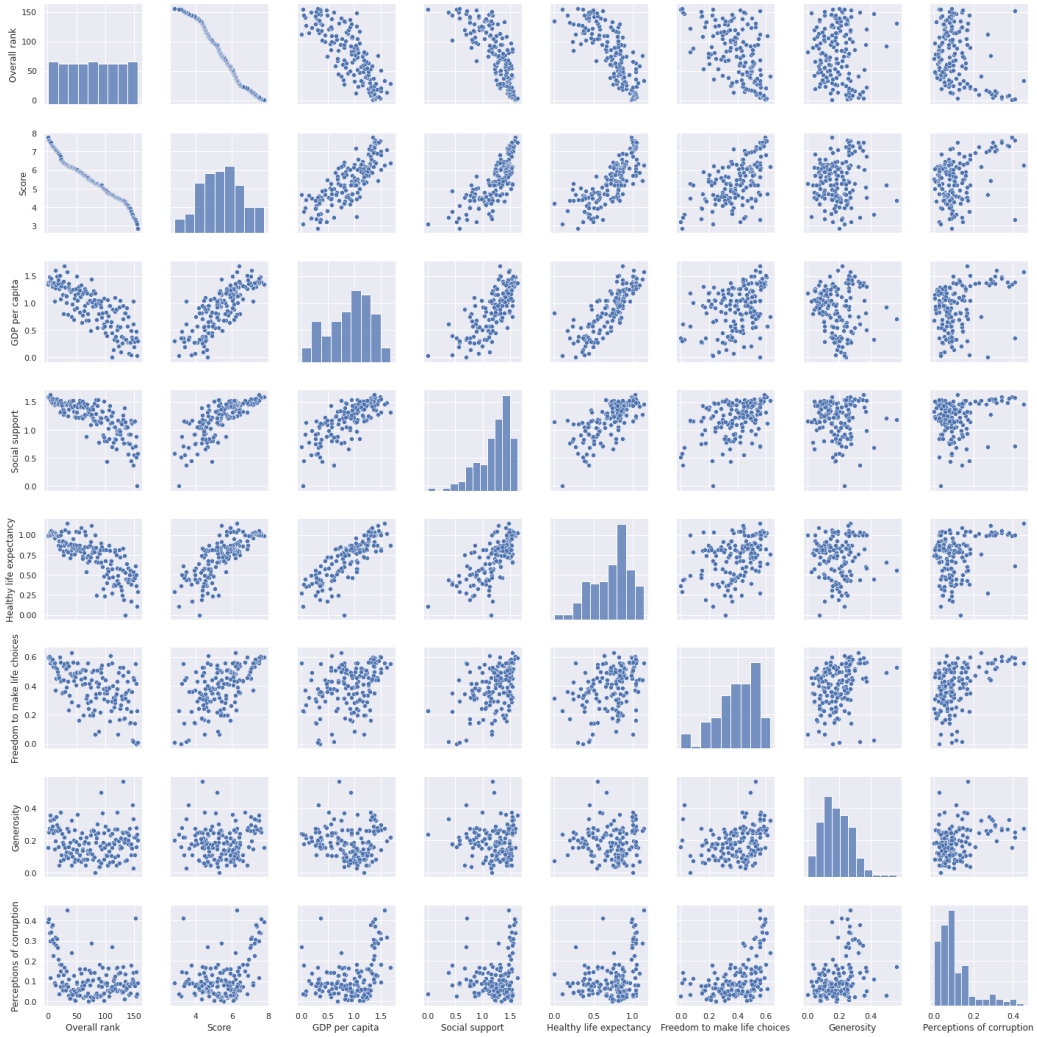
World Happiness Report (2019.csv)

https://www.kaggle.com/datasets/unsdsn/world-happiness?search=World+happiness+report&select=2019.csv

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2019, which ranks 156 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition as governments, organizations, and civil society increasingly uses happiness indicators to inform their policy-making decisions. Leading experts across fields – economics, psychology, survey analysis, national statistics, health, public policy, and more – describe how measurements of well-being can be used effectively to assess the progress of nations. The reports review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

Inference:

1. Pairwise plot



2. Correlation

For Pair of variable:

1. y = data['Overall rank'] # dependent variable

   x = data['Score'] # independent variable

   Here,
   R-squared: **0.978**, t: -83.346, p: 0.0
   As the R-squared is greater than 0.7 which is showing a high level of correlation.
   Also, there is a statistically significant difference in Overall rank and Score, t: -83.346, the null hypothesis is rejected,p: 0.0.
   Hence Samples are correlated.

2. y = data['Overall rank'] # dependent variable

   x = data[‘GDP per capita’] # independent variable

   Here,
   R-squared: **0.643**, t: -16.65, p: 0.0
   As the R-squared is less than 0.7 which is showing a low level of correlation.
   Also, there is a statistically significant difference in Overall rank and GDP per capita, t: -16.65, the null hypothesis is rejected,p: 0.0.
   Hence Samples are correlated.

3. y = data['Overall rank'] # dependent variable

   x = data[‘Social support’] # independent variable

   Here,
   R-squared: **0.59**, t:-14.856, p: 0.0
   As the R-squared is less than 0.7 which is showing a low level of correlation.
   Also, there is a statistically significant difference in Overall rank and Social support, t:-14.856, the null hypothesis is rejected,p: 0.0.
   Hence Samples are correlated.

4. y = data['Overall rank'] # dependent variable

   x = data['Healthy life expectancy’] # independent variable

   Here,
   R-squared: **0.62**, t:-15.85, p: 0.0
   As the R-squared is less than 0.7 which is showing a low level of correlation.
   Also, there is a statistically significant difference in Overall rank and Healthy life expectancy, t: 15.85, the null hypothesis is rejected,p: 0.0.
   Hence Samples are correlated.

5. y = data['Overall rank'] # dependent variable

   x = data['Freedom to make life choices’] # independent variable

   Here,
   R-squared: **0.3**, t:-8.1, p: 0.0
   As the R-squared is less than 0.7 which is showing a low level of correlation.
   Also, there is a statistically significant difference in Overall rank and Freedom to make life choices, t: 8.1, the null hypothesis is rejected,p: 0.0.
   Hence Samples are correlated.

6. y = data['Overall rank'] # dependent variable

   x = data['Generosity’] # independent variable

   Here,
   R-squared: **0.002**, t:-0.596, p: 0.0
   As the R-squared is less than 0.7 which is showing a low level of correlation.
   Also, there is a statistically significant difference in Overall rank and 'Generosity’, t: -0.596, the null hypothesis is accepted,p: 0.0.
   Hence Samples are uncorrelated.

   After performing log transformation, the R-squared: **0.003**, Hence Samples are uncorrelated.

7. y = data['Overall rank'] # dependent variable

   x = data[‘Perceptions of corruption’] # independent variable

Here,
R-squared: **0.124**, t:-4.666, p: 0.0
As the R-squared is less than 0.7 which is showing a low level of correlation.
Also, there is a statistically significant difference in Overall rank and Perceptions of corruption, t: -4.666, the null hypothesis is accepted,p: 0.0.
Hence Samples are correlated.

2. Determine if your regression or correlation statistics are 'significant'

Conditions:

1. If the p-value is less than the significance level (α=0.05):
   Decision: Reject the null hypothesis.
   Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

2. If the p-value is NOT less than the significance level (α=0.05)
   Decision: DO NOT REJECT the null hypothesis.
   Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Inference:

y = data['Overall rank'] # dependent variable
x = data['Score'] # independent variable
R-squared: **0.978**, p: 0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data['GDP per capita'] # independent variable
R-squared: **0.643**, p: 0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data['Social support'] # independent variable
R-squared: **0.59**, p: 0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data['Healthy life expectancy’] # independent variable
R-squared: **0.62**, p: 0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data['Freedom to make life choices’] # independent variable
R-squared: **0.3**, p:0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data['Generosity’] # independent variable
R-squared: **0.002**, p: 0.0
	Decision: DO NOT REJECT the null hypothesis.
	Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

y = data['Overall rank'] # dependent variable
x = data[Perceptions of corruption] # independent variable
R-squared: **0.124**, p: 0.0
	Decision: Reject the null hypothesis.
	Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."