Daffy Ducks
DSC 207: Mini-Project
Week 6 & 7

1 Mini-Project

Goal: To explore the 2016.csv dataset and present findings using data analysis and visualization techniques. In this mini-project, we will work with the dataset supplied and go through the data science process using the tools and techniques learned from Weeks 1 - 5.The aim is to potentially use these tools together to achieve the objective of data exploration.

1.1 High Level View [2 pts]

Describe the dataset in words (50 words). Look at the data samples and describe what they represent and how they could be useful in a variety of data science tasks.

1. The 2016.csv contains 13 fields, and 157 records.The dataset gives us 157 unique countries, that are split across 10 different regions. We are given some metrics for each country like happiness score, freedom, government corruption, family,and generosity. This data set could be useful for looking at how government corruption or freedom varies by region and country. It could also be useful for determining which of the metrics given is most influential on happiness score.

2 Data Exploration

In this step, you should explore what is present in the data and how the data is organized. You are expected to answer the following questions using the pandas library and markdown cells to describe your actions:

2.1 Preliminary Exploration

1. *Are there quality issues in the dataset (noisy, missing data, etc.)?*
    - Checking for inital quality issues int the dataset with df.info() where df is a pandas DataFrame object of our 2016.csv. Printing df.info gives us the following information:

column name	total non-null values	data type

column name	total non-null values	data type
country	157	object
region	157	object
happiness_score	157	float64
gdp_per_capita	157	float64
family	157	float64
life_expectancy	157	float64
freedom	157	float64
government_corruption	157	float64
generosity	157	float64
dystopia_residual	157	float64

Table 1 : Initial quality check of the 2016.csv dataset

2. *What will you need to do to clean and/or transform the raw data for analysis?*

   - Checking for null values (df.isnull().sum())
   - Dropping countries with a value 0 for an observation
   - The happiness_rank is an integer value that corresponds to the happiness_score, such that the max happiness is given a happiness_rank of 1, the min happiness is given a happiness_rank of n where n is the total number of records.Therefore, we need to reset the happiness rank to account for these dropped countries with an observation of 0.

3. *What are trends in the dataset using descriptive statistics (mean, median etc) and distribution of numerical data (eg. histograms)?*

   - In order to idenitify the global and statical trends in the data we can isolate the numerical columns categorical columns. By sepearting the numerical variables into a pandas dataframe we can following descriptive statistics for each numerical variable: mean, median, and standard deviation. We can compare mean and median to determine if the data is skewed.
   - By using histograms to visualize numerical data, we can see the distribution of the data. This will help us understand the data better and identify any skewness in these numerical variables.

2.2 Preliminary Exploration Tasks

You are expected to show a minimum of 2 preliminary exploration tasks that you performed with justification. Typically, preliminary exploration helps us in identifying specific objectives for data analysis tasks.

- Check for skewness in the data using histograms and descriptive statistics

1. In a distribution that is skewed right, the mean is greater than the median. While in a distribution that is skewed left, the mean is less than the median. For distrobutions that appear symmetric, the mean and median are roughly equal. By comparing the mean and median of the numerical variables in the 2016.csv dataset we can determine if a particular variable may contain outliers.

column name	mean	median
happiness_score	5.382185	5.314
gdp_per_capita	0.953880	0.982
family	0.793621	0.810
life_expectancy	0.557619	0.606
freedom	0.370994	0.397
government_corruption	0.137624	0.088
generosity	0.242635	0.222
dystopia_residual	2.325807	2.290

Table 2 : Mean and modes for numerical variables in the 2016.csv dataset

2. Histograms

- Check for null values or missing data in the dataset

Figure 1: Figure 1 : Histograms of numerical data in the 2016.csv dataset