Unit 5.4: Data Wrangling

1. The data sets that I used for this capstone project exist in one .csv file - obtained from Kickstarter(https://www.kaggle.com/kemical/kickstarter-projects#ks-projects-201801.csv)
2. In order to read in the .csv file, I first import pandas.
3. The .csv file ('ks-projects-201801.csv') is read in using pd.read_csv.
4. The dataset is checked for duplicate campaign ID's by using: df.duplicated, and examining if there are any duplicate rows. There are no duplicate rows.
5. The column names are examined first to see if there are any column names missing, or any names that should be changed. Additionally, I went through each column to understand the data housed within each column. I think about changing the column name of usd pledged to usd_pledged, but want to do some additional analysis before making that decision, as usd pledged and usd_pledged_real seem as though they are similar columns.
    a. **ID**: campaign ID
    b. **Name**: campaign name
    c. **Category**: project sub category
    d. **Main_category**: project main category
    e. **Currency**: what type of currency are the pledges
    f. **Deadline**: campaign deadline
    g. **Goal**: goal amount in the project currency
    h. **Launched**: date campaign launched
    i. **Pledged**: pledged amount in the project currency
    j. **State**: state of campaign (failed, successful, canceled)
    k. **Backers**: number of backers (the number of individuals who pledged funds to a campaign
    l. **Country**: country that campaign originated in
    m. **Usd pledged**: pledged amount in USD - this converts all pledged amounts that were in a different currency to USD for ease of analysis (originally done by kickstarter)
    n. **Usd_pledged_real**: pledged amount in USD - this converts all pledged amounts that were in a different currency to USD for ease of analysis (converted by fixer.io api)
    o. **Usd_goal_real**: goal amount in USD
6. I further investigate the dataset using .head() and .describe()
7. Next, I examined for null values using .isnull().sum() and found that only two columns contain null values. These columns are:
    a. Name (4)
    b. Usd Pledged (3797)

    The null values lead me to believe that the conversions completed by Kickstarter initially were inaccurate, which lead to the creation of usd_pledged_real using fixer.io api

instead. The usd_pledged_real column does not contain any null values. Because of this, I decided to drop the usd pledged column.
I next examined the null values from the name column, and found that the null names refer to four campaigns that were either cancelled or suspended. These null values are labeled as such (null (canceled), null(suspended)), and I decide to keep them this way as the rest of the data in regards to these campaign IDs is still available for analysis.

8. Outliers are identified by checking the data types of each column. Columns with data type ‘object’ are removed in order to determine z-scores over 3. After determining outliers, the shape of our data changes from 378661 rows to 375784 rows.