9.1 Final Project Step 2: Cleaning Your Data and Exploratory Data Analysis

Rachel Nelson

8/1/2020

```
# load the data
ks_df <- read.csv("C:/Users/Rachel/Desktop/College/DSC520/dsc520/data/ks-projects-201801.csv")
```

1. Data importing and cleaning steps are explained in the text and in the Github exercises. (Tell me why you are doing the data cleaning activities that you perform). Follow a logical process.

I am cleaning the data set to prepare it for analysis.

Check for missing columns

```
# Check for Missing Columns
names(ks_df)
```

```
##  [1] "ID"               "name"             "category"         "main_category"
##  [5] "currency"         "deadline"         "goal"             "launched"
##  [9] "pledged"          "state"            "backers"          "country"
## [13] "usd.pledged"      "usd_pledged_real" "usd_goal_real"
```

```
ks_df$rowid <- paste(ks_df$ID, "-", ks_df$round)
length(unique(ks_df$rowid))
```

```
## [1] 378661
```

```
length(ks_df$rowid)
```

```
## [1] 378661
```

Here I confirmed that all rows have a unique ID. I also reviewed the data to ensure all the data I needed was contained within the data set.

Check variables names

```
# checks variable names and replace with new name
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
ks_df <- rename(ks_df, usd_pledged = usd.pledged)
```

Here I renamed the variable usd.pledged to usd_pledged to align the naming conventions of all of my headers, since the rest of the headers uses underscores instead of periods for spaces.

Check missing observations

```
# checks for missing values in observations
colMeans(is.na(ks_df))
```

```
##               ID             name         category    main_category
##       0.00000000       0.00000000       0.00000000       0.00000000
##         currency         deadline             goal         launched
##       0.00000000       0.00000000       0.00000000       0.00000000
##          pledged            state          backers          country
##       0.00000000       0.00000000       0.00000000       0.00000000
##      usd_pledged usd_pledged_real    usd_goal_real            rowid
##       0.01002744       0.00000000       0.00000000       0.00000000
```

```
# removes column from data set
ks_df = subset(ks_df, select = -c(usd_pledged) )
```

Here I am looking for missing values. There is a small amount of data in the usd_pledged with missing values. If I wanted to cleanse the data set, I could remove these values, but for now, I want to keep it in mind since there are zero missing values from usd_pledged_real, which is a column giving the same information, but the conversion to USD was done from the fixer.io api instead of done by kickstarter. Instead of removing the rows with the missing data, I am going to remove the column from the data set since it is a duplicate column.

usd_pledged: conversion in US dollars of the pledged column (conversion done by kickstarter). usd pledge real: conversion in US dollars of the pledged column (conversion from Fixer.io API).

Check variable classification

```
# checks attributes of data frame
str(ks_df)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID              : int  1000002330 1000003930 1000004038 1000007540 1000011046 1000014025 10000234
##  $ name            : Factor w/ 375765 levels "","\177Not Twins - New EP! \"The View from Down Here\"
##  $ category        : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 56 124 59 42 114 40 ...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7 11 7 8 8 8 5 7 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14 14 14 14 14 14 14 14 14 ...
##  $ deadline        : Factor w/ 3164 levels "2009-05-03","2009-05-16",..: 2288 3042 1333 1017 2247 24
##  $ goal            : num  1000 30000 45000 5000 19500 50000 1000 25000 125000 65000 ...
```

```
##  $ launched        : Factor w/ 378089 levels "1970-01-01 01:00:00",..: 243292 361975 80409 46557 235
##  $ pledged         : num  0 2421 220 1 1283 ...
##  $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2 2 2 1 4 4 2 1 1 ...
##  $ backers         : int  0 15 3 1 14 224 16 40 58 43 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 23 23 23 23 23 23 23 23 ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...
##  $ rowid           : chr  "1000002330 - " "1000003930 - " "1000004038 - " "1000007540 - " ...
```

Checking the variable classification is the step used to make sure the data is the right datatype for analysis.

Check duplicate rows

```
# Checking if one row is identical to another
distinctdata <- distinct(ks_df)
nrow(ks_df)
```

```
## [1] 378661
```

```
nrow(distinctdata)
```

```
## [1] 378661
```

Checking for duplicate rows within the data. None were found. If duplicate rows are found, the duplicate should be extracted from the dataset.

Change dates from factors to date

```
ks_df <- transform(ks_df, deadline = as.Date(deadline), launched = as.Date(launched))
```

Changes the data type of deadline and launched to date.

2. With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

```
head(ks_df)
```

```
##           ID                                                  name
## 1 1000002330                        The Songs of Adelaide & Abullah
## 2 1000003930             Greeting From Earth: ZGAC Arts Capsule For ET
## 3 1000004038                                          Where is Hank?
## 4 1000007540          ToshiCapital Rekordz Needs Help to Complete Album
## 5 1000011046 Community Film Project: The Art of Neighborhood Filmmaking
## 6 1000014025                                    Monarch Espresso Bar
##         category main_category currency   deadline  goal   launched pledged
## 1         Poetry    Publishing      GBP 2015-10-09  1000 2015-08-11       0
## 2 Narrative Film  Film & Video      USD 2017-11-01 30000 2017-09-02    2421
## 3 Narrative Film  Film & Video      USD 2013-02-26 45000 2013-01-12     220
## 4          Music         Music      USD 2012-04-16  5000 2012-03-17       1
## 5   Film & Video  Film & Video      USD 2015-08-29 19500 2015-07-04    1283
## 6    Restaurants          Food      USD 2016-04-01 50000 2016-02-26   52375
```

```
##        state backers country usd_pledged_real usd_goal_real         rowid
## 1     failed       0      GB                0       1533.95 1000002330 -
## 2     failed      15      US             2421      30000.00 1000003930 -
## 3     failed       3      US              220      45000.00 1000004038 -
## 4     failed       1      US                1       5000.00 1000007540 -
## 5   canceled      14      US             1283      19500.00 1000011046 -
## 6 successful     224      US            52375      50000.00 1000014025 -
```

3. What do you not know how to do right now that you need to learn to import and cleanup your dataset?

I need to figure out if and how the factor/category data needs to be changed to numerical data. I also had to change dates from factors to date data types.

4. Discuss how you plan to uncover new information in the data that is not self-evident.

I plan to run both correlation and unsupervised learning models on the data to see if I can uncover any new information that is not self-evident.

5. What are different ways you could look at this data to answer the questions you want to answer?

Yes, the questions I want to answer can be viewed though looking at bar charts, frequency plots and statistical models. * Are there certain types/category of campaigns that are more successful? * How much money should you ask for? * Is there a time period for the campaign that works better than others? * What is the average contribution of a backer? * Is there a better time of year to launch a campaign?

6. Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I might create a new variable for % successful by taking the pledged and dividing it by the goal.

7. How could you summarize your data to answer key questions?

This ties into the different ways I can look at the data set. Charts and visualizations are a great way to summarize the data and answer key questions.

8. What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

Bar charts, box plots and scatter charts will help illustrate findings to my questions.

My Research Questions: * Are there certain types/category of campaigns that are more successful? * How much money should you ask for? * Is there a time period for the campaign that works better than others? * What is the average contribution of a backer? * Is there a better time of year to launch a campaign?

9. What do you not know how to do right now that you need to learn to answer your questions?

This still ties in to question #3, where I need to figure out if the factor/categy data needs to be changed to numerical data and if so, how I go about doing that.

10. Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Yes, I plan to see if there are any supervised (like decision tree or random forest) models and unsupervised (clustering) that can help make sense of what is funded verses unfunded.