Universitat Oberta de Catalunya

Master Ciencia de Datos

Tratamiento de datos: *Kickstarter Projects*.

Irene López Ruiz

Javier Samir Rey

Índice

1 Descripción del dataset

Kickstarter (https://www.kickstarter.com) es una plataforma de mecenazgo para proyectos creativos (*Kickstarter e impuestos — Kickstarter*, s. f.). Los dueños del proyecto eligen una fecha límite y un mínimo objetivo de fondos a recaudar y, si el objetivo elegido no es recolectado en el plazo, no se perciben fondos gracias a un contrato de garantía (*Kickstarter - Wikipedia, la enciclopedia libre*, s. f.).

Como es lógico, no todos los proyectos de Kickstarter consiguen recaudar los fondos que necesitan, mientras que otros rebasan por grandes cantidades la suma mínima que requerían sus autores. Esto plantea la necesidad de esclarecer las características de los proyectos que consiguen recaudar el dinero necesario y las diferencias que éstos tienen con los proyectos que son finalmente cancelados. Con un estudio de este calibre se podrán realizar estrategias de lanzamiento de proyectos para que los autores de estos tengan una mayor probabilidad de éxito, beneficiando así tanto a los proyectos como a la propia plataforma.

Es por esto por lo que se ha decidido realizar un proyecto de análisis de datos sobre un conjunto de datos, provisto en la plataforma Kaggle (https://www.kaggle.com), sobre la financiación de diferentes proyectos de Kickstarter que permita discriminar los puntos clave de los proyectos que más éxito tienen, tanto en sí consiguen los fondos que planteaban como objetivo como en la cantidad de dinero que consiguen recaudar por encima de dicho objetivo. El conjunto de datos tratado ha sido creado por el usuario de Kaggle Mickaël Mouillé (https://www.kaggle.com/kemical), y la documentación del mismo puede encontrarse en https://www.kaggle.com/kemical/kickstarter-projects?select=ks-projects-201801.csv

Este conjunto de datos se compone de dos ficheros CSV, siendo uno de ellos (*ks-projects-201612.csv*) un subconjunto de los registros del otro (*ks-projects-201801.csv*), por lo que solo se tratará este último fichero, más completo. Este *dataset* cuenta con un total de 378661 proyectos, descritos con 15 variables diferentes:

- **ID**: identificador numérico y único del proyecto en la plataforma Kickstarter.
- **Name**: Nombre del proyecto.
- **Category**: categoría de clasificación del proyecto, por ejemplo, poesía.
- **main_category**: categoría principal para la clasificación del proyecto en clases más generales, por ejemplo, arte.
- **Currency**: acrónimo de la moneda de la financiación del proyecto, por ejemplo, USD.
- **Deadline**: fecha de finalización de la campaña de financiación del proyecto.
- **Goal**: objetivo de financiación propuesto por los autores del proyecto. Cantidad indicada en la moneda especificada por el campo *currency*.
- **Launched**: fecha y hora de lanzamiento de la campaña de financiación del proyecto.
- **Pledged**: cantidad recaudada por el proyecto, expresada en la moneda especificada por el campo *currency*.
- **State**: estado del proyecto. Posee varias categorías posibles, como cancelado, fallido, o existoso.
- **Backers**: número de patrocinadores o mecenas del proyecto.
- **Country**: país de origen del proyecto, descrito mediante el acrónimo de dos letras, por ejemplo, US.
- **usd_pledged**: cantidad recaudada por el proyecto, expresada en dólares americanos (USD), siendo la conversión entre la moneda original realizada por Kickstarter.

- **usd_pledged_real**: cantidad recaudada por el proyecto, expresada en dólares americanos (USD), siendo la conversión entre la moneda original realizada por la plataforma Fixer (https://fixer.io).
- **usd_goal_real**: objetivo de financiación propuesto por los autores del proyecto. Cantidad indicada en dólares americanos (USD) siendo la conversión entre la moneda original realizada por la plataforma Fixer (https://fixer.io).

El presente documento supone una memoria con la descripción de los pasos y conclusiones extraídas del preprocesado y análisis de estos datos realizado con el código que se adjunta en este proyecto.

2 Lectura y selección de los datos de interés

En el código adjunto a este proyecto se ha realizado, tras la carga del fichero, un resumen de todas las variables que posee el conjunto de datos donde, entre otros, se muestra se muestra la codificación en R de cada variable, que deberá ser subsanada para que considere como variables categóricas a todos aquellos campos que toman para cada registro un valor entre una serie de clases concretas; y también para las variables con fechas (*launched* y *deadline*), que deberán codificarse como datos temporales. Se ha comprobado también que no existen duplicados de registros que deban borrarse para no considerar el mismo proyecto más de una vez. Tras esta correcta lectura de los datos del fichero se procede a preprocesar los datos.

En primer lugar, se procede a una selección de las variables a analizar. Recordemos que se dispone de los datos objetivo y recaudación sobre la financiación del proyecto, tanto en dólares americanos como en la moneda original. Esta información es redundante y, considerando que los datos en dólares americanos resultan una estandarización que facilita la comparación entre diferentes proyectos, las variables que referencian la financiación en la moneda original serán eliminadas. Además, la conversión a dólares realizada por Kickstarter (*usd pledged*) no resulta del todo satisfactoria ya que, como se aprecia en el resumen de las variables, presenta datos perdidos a pesar de que la variable original *pledged* no posee datos perdidos. La conversión realizada por Fixer (*usd_pledged_real*) resulta más completa y satisfactoria, luego será esta variable la que represente la cantidad recuadada en dólares.

Por tanto, se seleccionarán todas las variables excepto *goal*, *pledged*, *currency* y *usd pledged*.

3 Limpieza de los datos

Continuando con una limpieza de datos más exhaustiva se han realizado acciones de reducción del volumen de datos, conversión y creación de nuevas variables, tratamiento de datos perdidos y análisis de los valores extremos. A continuación, se describen estos procesos.

3.1 Reducción de casos

A parte de la selección de variables realizada y comentada anteriormente, se realiza una reducción del número de registros, bastante elevado. En concreto, teniendo en cuenta que uno de los intereses del análisis de estos datos es conocer el estado de los mismos para poder distinguir las características clave de los proyectos exitosos, los proyectos con estado indefinido (valor *undefined* en el campo *state*) serán eliminados.

3.2 Conversión y creación de nuevas variables

Otro proceso habitual en el preprocesado o limpieza de datos consiste en la conversión de variables y la creación de nuevos campos que permitan un análisis posterior más eficiente y fácil de interpretar.

Con este propósito se han realizado dos acciones: en primer lugar, se poseen dos campos, *launched* y *deadline*, que representan la fecha de lanzamiento y de final de la campaña de recaudación de los proyectos, respectivamente. Sin embargo, más que las fechas concretas de estos periodos, interesa una localización más general del lanzamiento del proyecto en el tiempo, como pueda ser el año en que este comenzó. Además, la fecha concreta de finalización de la recaudación puede ser sustituida por una variable más explicativa que represente la duración, en días, de la campaña de financiación que, sin duda, influirá en la cantidad recaudada. Por tanto, estas dos variables se sustituirán por dos nuevas variables *launch_year*, con el año de lanzamiento del proyecto, y *days_of_campaign*, con el número de días para la financiación desde el lanzamiento del proyecto.

En segundo lugar, se ha creado una nueva variable con la diferencia entre el dinero recaudado y el dinero objetivo del proyecto. En este caso, esta nueva variable no sustituye a ninguna otra, si no que complementa los datos. Esto se debe a que facilitará el análisis de los datos para concluir qué proyectos resultan más exitosos y eficientes en sus campañas de recaudación y consiguen no solo alcanzar su financiación mínima requerida, si no sobrepasarla en gran cantidad. De forma similar, permitirá cuantificar la desviación de los proyectos que más se alejan de conseguir su objetivo de financiación para analizar cuáles son las características comunes a estos proyectos. Sin embargo, no solo esta diferencia resulta importante a la hora de realizar los estudios de viabilidad del proyecto, también influirá en el éxito o fracaso de los proyectos la cantidad inicial que reclaman sus autores puesto que, si esta es muy elevada podría disuadir a los posibles inversores por ser un proyecto inviable. Por tanto, se creará la variable *diff_pledge_real* con la diferencia entre dinero recaudado y dinero solicitado por los proyectos, pero se mantendrán estas dos últimas variables en el conjunto, por resultar también de interés.

3.3 Datos perdidos

En el resumen de las variables realizado durante la carga de los datos, entre otros, se apreciaban la cantidad de datos perdidos que posee cada variable. La única variable con datos perdidos, como ya se comentó, era *usd_pledged*, que fue descartada durante la etapa de selección de variables. Sin embargo, no es cierto que no se presenten más valores perdidos ya que, en la variable *country* se encuentran valores 'N, 0"', que no se corresponden con un país si no que codifican valores perdidos para el país de origen del proyecto. Si bien resulta imposible y carente de sentido realizar una imputación de los países, sí que se puede proceder a codificarlos de una forma más representativa con la palabra 'unknown'.

3.4 Valores extremos

Se ha realizado un análisis de los valores extremos de las variables numéricas del conjunto de datos: *backers*, *usd_goal_real*, *usd_pledged_real*, *days_of_campaign*. Se encuentra una gran multitud de valores atípicamente altos en comparación con el resto de valores suelen ser cercanos a cero. Estos valores atípicos modifican notablemente las estadísticas y distribuciones de las variables.

Además, se realiza una comparación gráfica y numérica de la distribución de cada una de estas variables considerando los valores atípicos y una corrección de los mismos. Las variables sin

valores atípicos resultan más contundentes y poseen una distribución con más sentido, lo que podría indicar que todos estos valores atípicos suponen errores de codificación que no representan realmente los valores correctos. Sin embargo, basta buscar algunos de los proyectos que presentan estos valores atípicos en Kickstarter para darse cuenta de que no se trata de errores en la creación del *dataset*, si no que realmente estos proyectos cuentan con valores tan altos en su número de patrocinadores y respecto a su financiación.

Por tanto, estos valores extremos no serán tratados ya que se asume, y será parte del análisis, que hay proyectos que no consiguen nada o muy poca financiación mientras que existen proyectos que sobrepasan con mucha diferencia los objetivos de financiación asumidos.

4 Análisis de los datos

4.1 Selección de los grupos de datos que se quieren analizar/comparar

Se van a realizar análisis de prueba de hipótesis entre el país de estados unidos y el país de Gran Bretaña con base en el valor de meta de recaudo en dólares, es importante anotar que estados unidos cuenta con muchas más solicitudes que también tienen monedas diferentes, aunque los valores son estandarizados a dólares. Ver gráfica 3 (meta solicitada USD y GB).

4.2 Comprobación de la normalidad y homogeneidad de la varianza

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, se utilizará la prueba de normalidad de Anderson-Darling. Para cada prueba se obtiene un p-valor superior al nivel de significación prefijado **α= 0,05** para una confianza de un 95%. Si esto se cumple, entonces se considera que variable sigue una distribución normal esto es particularmente útil porque la distribución normal tiene ciertas propiedades matemáticas de simetría que permite estudiar los fenómenos que modela de acuerdo con unos patrones ya documentados.

Recordemos que, gracias a las propiedades simétricas, Cuando la distribución es normal el 68% de las observaciones se encuentran entre ± una desviación estándar de la media, el 95% entre dos desviaciones estándar y el 99,7% entre tres. Es una regla empírica.

Variables que no siguen una distribución normal:

nid, goal, npledged, backers, nusd_pledged, usd_pledged_real, usd_goal_real, launch_year, days_of_campaign, diff_pledge_real.

Homogenidad

En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los tipos de categorias de crowdfunding. En la siguiente prueba, la hipótesis nula consiste en que ambas varianzas son iguales.

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  usd_goal_real by main_category
## Fligner-Killeen:med chi-squared = 63264, df = 14, p-value < 2.2e-16
```

Dado que se obtiene un p-valor inferior a 0.05, rechazamos la hipótesis de que las varianzas de que las muestras por categorias son homogéneas.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

Se procede a realizar un análisis decorrelación entre las distintas variables (numéricas) para determinar cuáles de ellas ejercen una mayor influencia sobre el valor de meta de la campaña en dólares (**usd_goal_real**). Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
##                     estimate      p-value
## id                0.002455738 1.326970e-01
## goal              0.215731884 0.000000e+00
## pledged           0.019230143 5.256478e-32
## backers          -0.002211068 1.758166e-01
## usd_pledged       0.019958529 2.400739e-34
## usd_pledged_real  0.017152048 8.432359e-26
## launch_year      -0.102129201 0.000000e+00
## days_of_campaign  1.000000000 0.000000e+00
```

Con base en los resultados, se puede identificar cuáles son las variables más correlacionadas con el valor de meta en dólares en función de su aporte entre los valores -1 y +1. Teniendo esto en cuenta. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

Teniendo en cuenta variable con mayor estimado, la variable más relevante en la obtención de fondos para el crowdfunding es la variable: **days_of_campaign**, dado que es el valor solicitado.

4.4 Aplicación de pruebas estadísticas para comparar los grupos de datos

Esta prueba estadística se usa para realizar un contraste de hipótesis sobre dos muestras (en nuestro caso son los datos completos de la población) para determinar si la meta de recaudo del proyecto es superior dependiendo si el país es US o GB. Con 2 muestras: la primera de ellas corresponde a los valores de meta para US, la segunda, para los valores de meta de GB. Dado que la muestra es un n>30 podemos usar la prueba.

Planteado dos hipótesis:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 < 0$$

Donde **μ1** es la media de la población de la que se extrae la primera muestra y **μ2** es la media de la población de la que extrae la segunda. Así, tomaremos **α=** 0,05, para tener una confianza de 95%.

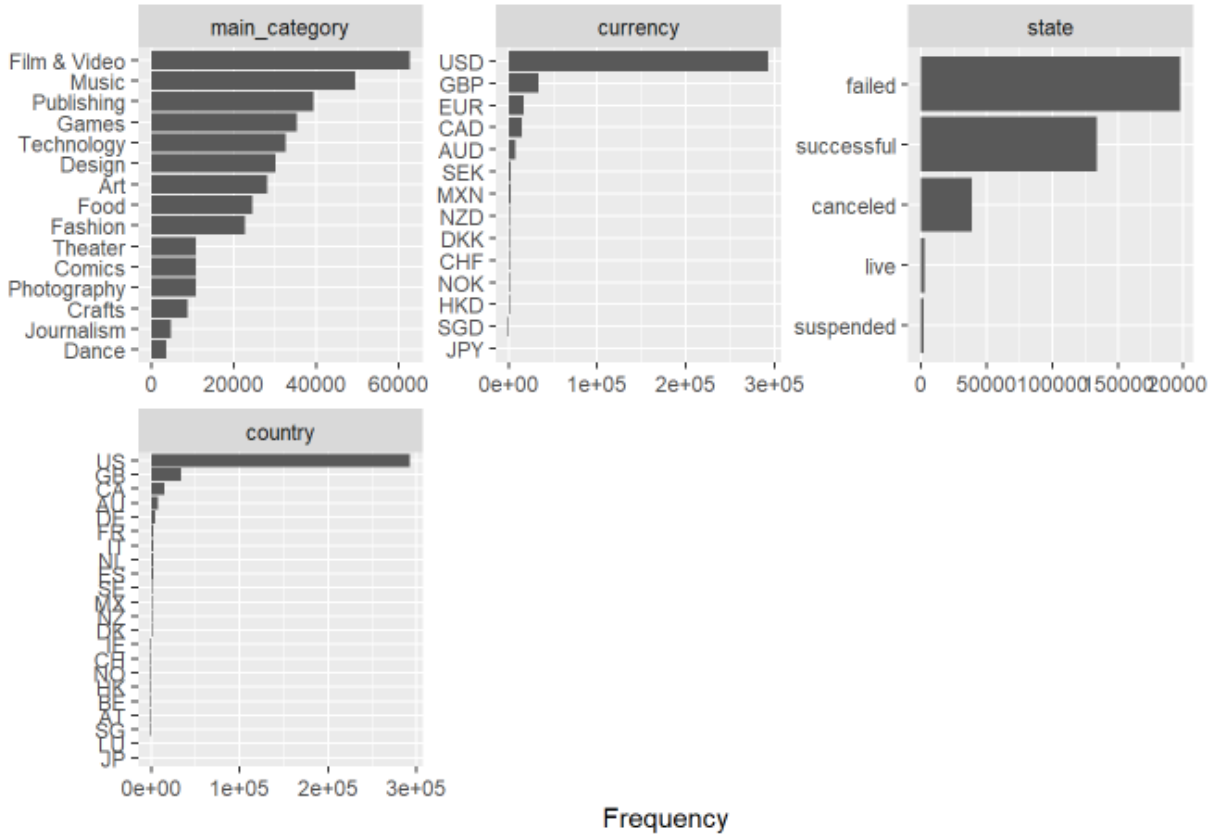
```
        One Sample t-test

data:  goal_usd_us
t = 21.492, df = 292626, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 40019.11 48050.83
sample estimates:
mean of x
 44034.97
```

Puesto que obtenemos un **p-valor** menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, las metas de recaudo en estados unidos son

diferentes a la media de las metas de recaudo de Gran Bretaña.

5 Representación de los resultados a partir de tablas y gráficas

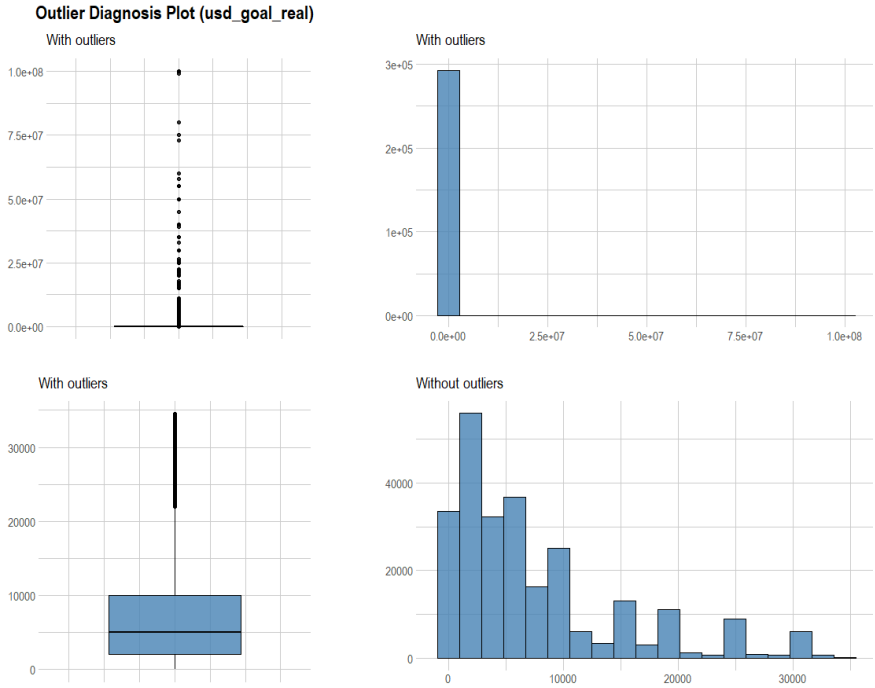
5.1 Variables discretas



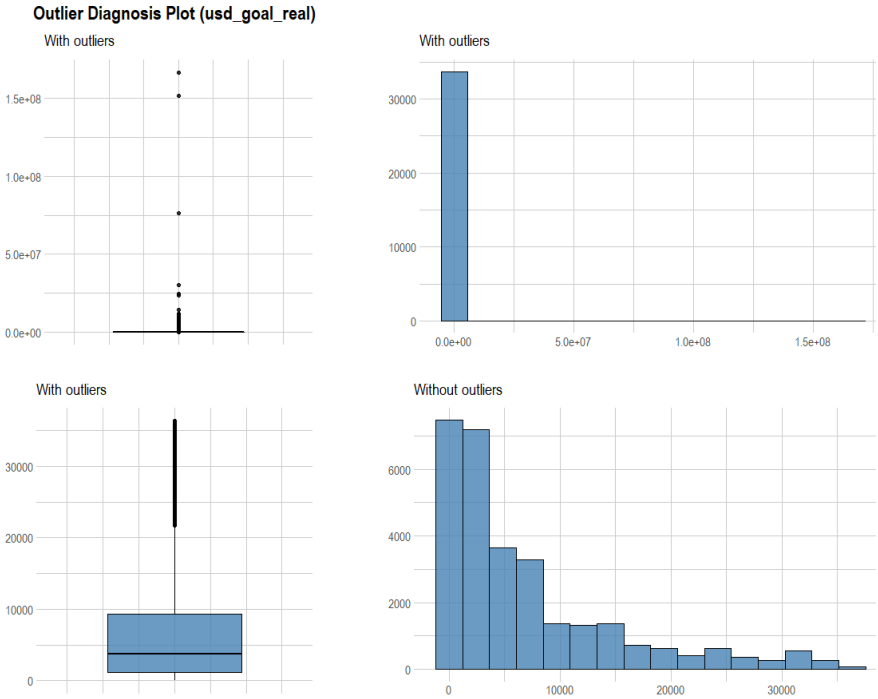
5.2 Detección de valores anómalos

variables <chr>	outliers_cnt <int>	outliers_ratio <dbl>	outliers_mean <dbl>	with_mean <dbl>	without_mean <dbl>
backers	44713	11.92779	722.38823	106.68728	23.30168
usd_goal_real	44700	11.92432	327236.95622	45863.17875	7768.75340
usd_pledged_real	50558	13.48702	58217.57049	9120.80292	1466.81448
days_of_campaign	76235	20.33671	46.37488	34.49768	31.46562

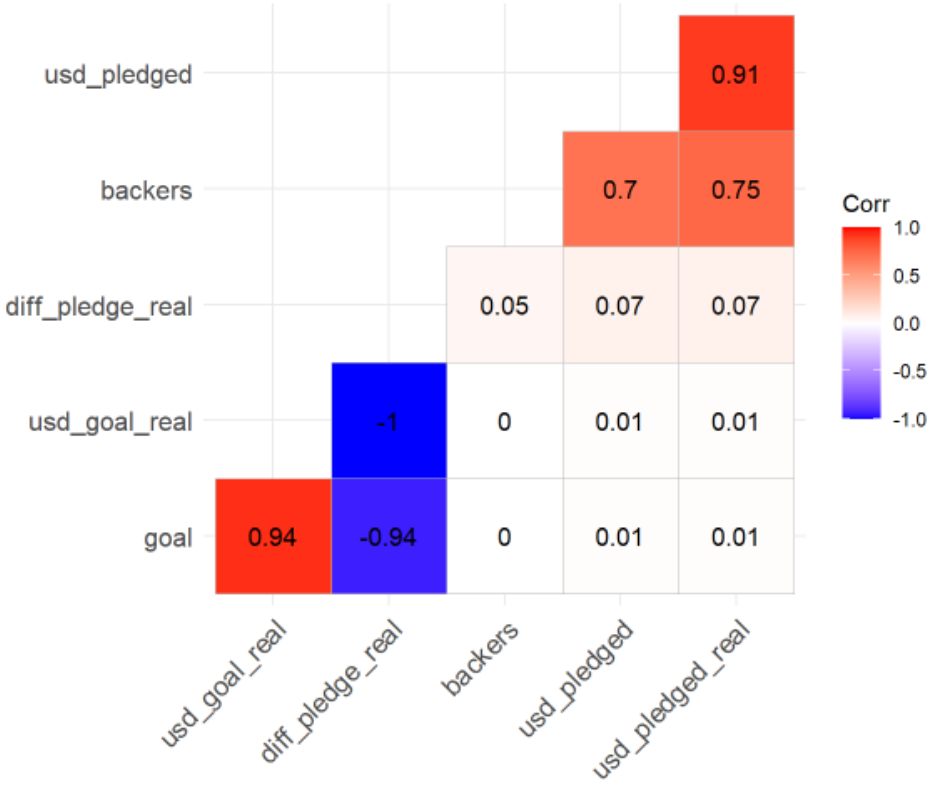
5.3 Meta solicitada US con y sin valores extremos



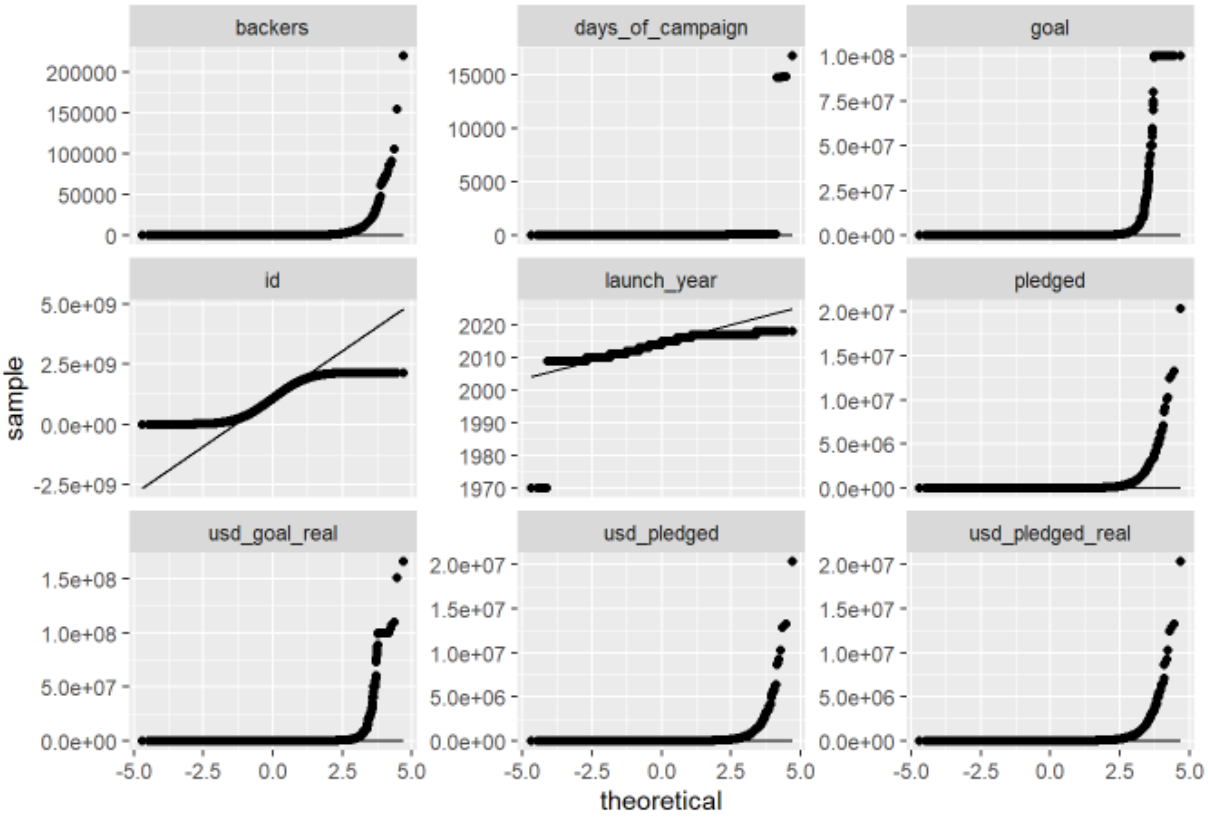
5.4 Meta solicitada GB con y sin valores extremos



5.5 Correlación de variables numéricas



5.6 Pruebas de normalidad



6 Conclusiones

Hemos podido apreciar que el conjunto de datos de **Kickstarter** tiene una historia y cantidad relevante de registros que permiten realizar diversos análisis y de esta manera tener conclusiones relevantes acerca de cómo proyectos de fondeo comunitario se comportan.

Se evaluó la normalidad de los datos encontrando que la mayoría no tienen una distribución normal, presentando un sesgo hacia la izquierda (ver imágenes de apoyo), sin embargo, al remover outlier se puede notar que las medias entre US y GB son similares.

También los datos muestran evidencia de que las categorias y la meta de recaudo no son homogéneas entre categorias de acuerdo con la prueba realizada.

De acuerdo con la validación del aporte de las variables según la meta de recaudo el modelo evidencia que la variable que mas aporta y correlacionada es: la cantidad de días que dura la campaña.

Por último, según la prueba de contraste de hipótesis la media de la cantidad de dinero solicitado de US tiene un diferencia significativa con la cantidad de dinero solicitado en Gran Bretaña.

Para un siguiente análisis sería prometedor combinar estos datos con otras fuentes de información como valores demográficos y detalles concretos de los proyectos financiados.

7 Bibliografía

Kickstarter - Wikipedia, la enciclopedia libre. (s. f.). Recuperado 30 de mayo de 2021, de https://es.wikipedia.org/wiki/Kickstarter

Kickstarter e impuestos — Kickstarter. (s. f.). Recuperado 30 de mayo de 2021, de https://www.kickstarter.com/help/taxes?lang=es

ANEXO: Contribuciones

Contribuciones	Firma
Investigación previa	ILR JSR
Redacción de las respuestas	ILR JSR
Desarrollo de código	ILR JSR