Co?mo crear funciones en Python-Removiendo outliers

August 5, 2017

In [2]:
```
import pandas as pd
import numpy as np
```

In [16]:
```
dfnum=pd.read_csv('funciones.csv')
del dfnum['Unnamed: 0']
```

In [17]:
```
low = .005
high = .995
quant_df = dfnum.quantile([low, high])
filt_df = dfnum.apply(lambda x: x[(x>=quant_df.loc[low,x.name]) & (x <= quant_df.loc[l
col_names=['total_minutes', 'min_dif','found_rate_pickers', 'picking_speed_pickers','a
           'rating_pickers','found_rate_drivers', 'picking_speed_drivers','accepted_ra
           'quantity_UN','found_rate_UN']
filt_df = filt_df.dropna(axis=0,how='any',thresh=None,subset=col_names)
print(filt_df.shape)
round(100*(filt_df.isnull().sum(axis=0)/filt_df.shape[0]),1)
```

```
(6077, 14)
```

Out[17]:
```
total_minutes             0.0
min_dif                   0.0
found_rate_pickers        0.0
picking_speed_pickers     0.0
accepted_rate_pickers     0.0
rating_pickers            0.0
found_rate_drivers        0.0
picking_speed_drivers     0.0
accepted_rate_drivers     0.0
rating_drivers            0.0
quantity_Kg              38.4
found_rate_Kg            38.5
quantity_UN               0.0
found_rate_UN             0.0
dtype: float64
```

In [18]:
```
#Creamos la función que extrae el % de valores inferiores y exteriores de las columna
def get_away_outliers(df,low=0.005,high=0.995,column_names=None,how='any'):
```

```
        quant_df = df.quantile([low, high])
        filt_df = df.apply(lambda x: x[(x>=quant_df.loc[low,x.name]) & (x <= quant_df.loc
        if column_names==None:
            if how=='any':
                filt_df = filt_df.dropna(axis=0,how='any',subset=column_names)
            else:
                filt_df = filt_df.dropna(axis=0,how='all',subset=column_names)
        else:
            if how=='any':
                filt_df = filt_df[column_names].dropna(axis=0,how='any',subset=column_name
            else:
                filt_df = filt_df[column_names].dropna(axis=0,how='all',subset=column_name
        return filt_df

    #Fuentes:
    #https://stackoverflow.com/questions/35827863/remove-outliers-in-pandas-dataframe-usi
    #https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.dropna.html
```

In [19]:
```
col_names=['total_minutes', 'min_dif','found_rate_pickers', 'picking_speed_pickers','a
           'rating_pickers','found_rate_drivers', 'picking_speed_drivers','accepted_ra
           'quantity_UN','found_rate_UN']
filt_df=get_away_outliers(df=dfnum,column_names=col_names)
round(100*(filt_df.isnull().sum(axis=0)/filt_df.shape[0]),1)
```

Out[19]:
```
total_minutes            0.0
min_dif                  0.0
found_rate_pickers       0.0
picking_speed_pickers    0.0
accepted_rate_pickers    0.0
rating_pickers           0.0
found_rate_drivers       0.0
picking_speed_drivers    0.0
accepted_rate_drivers    0.0
rating_drivers           0.0
quantity_Kg             38.4
found_rate_Kg           38.5
quantity_UN              0.0
found_rate_UN            0.0
dtype: float64
```

In [1]:
```
#Creamos la función que genera columnas sin  el % de valores inferiores y exteriores d
#queramos de un df y elimina posteriormente aquellas filas, en donde también podemos d
def get_away_outliers(df,low=0.005,high=0.995,col_outlier=None,col_na=None,how='any'):
    quant_df = df.quantile([low, high])
    if col_outlier==None:
        filt_df = df.apply(lambda x: x[(x>=quant_df.loc[low,x.name]) & (x <= quant_df.l
        if how=='any':
            df = filt_df.dropna(axis=0,how='any',subset=col_na)
```

```
        else:
            df = filt_df.dropna(axis=0,how='all',subset=col_na)
    else:
        filt_df = df[col_outlier].apply(lambda x: x[(x>=quant_df.loc[low,x.name]) & (x
        df.loc[:,col_outlier]=filt_df
        if how=='any':
            filt_df = filt_df.dropna(axis=0,how='any',subset=col_na)
            df=df.loc[filt_df.index,:]
        else:
            filt_df = filt_df.dropna(axis=0,how='all',subset=col_na)
            df=df.loc[filt_df.index,:]
    return df

#Fuentes:
#https://stackoverflow.com/questions/35827863/remove-outliers-in-pandas-dataframe-using
#https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.dropna.html
```