

```
import pandas as pd
import numpy as np
#pd.set_option('max_columns', 120)
#pd.set_option('max_colwidth', 5000)

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
plt.rcParams['figure.figsize'] = (12,8)
```

```
#-----
# THIS DATA IS NOT CONTINUED FROM THE PREV LAB FOR SIMPLICITY
#-----
```

```
filtered_loans = pd.read_csv('https://github.com/kaopanboonyuen/2110446_DataSci
#drop_cols = ['last_credit_pull_d','title']
#filtered_loans = filtered_loans.drop(drop_cols,axis=1)
print(filtered_loans.shape)
filtered_loans.head()
```

```
null_counts = filtered_loans.isnull().sum()
print("Number of null values in each column:\n{}".format(null_counts))
```

```
filtered_loans
```

```
filtered_loans.mean()
```

```
filtered_loans.fillna(filtered_loans.mean(), inplace=True)
```

```
null_counts = filtered_loans.isnull().sum()
print("Number of null values in each column:\n{}".format(null_counts))
```

```
# Imputation of missing values for categories in pandas using 'mode'
filtered_loans_v1 = filtered_loans.copy()
```

```
# For 'mode', there can be many outputs, so we pic the first one.
filtered_loans_v1.fillna(filtered_loans_v1.mode().iloc[0], inplace=True)
```

```
null_counts = filtered_loans_v1.isnull().sum()
print("Number of null values in each column:\n{}".format(null_counts))
```

✓ SimpleImputer (Alternative Method)

Imputation transformer for completing missing values.

Credit: <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>

```
filtered_loans = pd.read_csv('https://github.com/kaopanboonyuen/2110446_DataSc
null_counts = filtered_loans.isnull().sum()
print("Number of null values in each column:\n{}".format(null_counts))
```

```
# 1) select only numeric columns
filtered_loans_num = filtered_loans[['revol_util', 'pub_rec_bankruptcies']]
```

```
from sklearn.impute import SimpleImputer

num_imp=SimpleImputer(missing_values=np.NaN, strategy='mean')
filtered_loans[['revol_util', 'pub_rec_bankruptcies']]=pd.DataFrame(num_imp.fit_
filtered_loans.isnull().sum())
```

```
# 2) select only categorical columns
filtered_loans_cat = filtered_loans[['emp_length']]
```

```
cat_imp=SimpleImputer(missing_values=np.NaN, strategy='most_frequent')
filtered_loans['emp_length']=pd.DataFrame(cat_imp.fit_transform(filtered_loans.
filtered_loans.isnull().sum())
```

