LAB #1

DATA PREPROCESSING

(EXPLORATION, CLEANSING, TRANSFORM, AND FEATURE SELECTION)

LAB#1

#1.1: Data Preparation

- Data Exploration
- Data Cleansing
- Data Transform

#1.2: Feature Selection

- Numerical Feature Selection (Correlation)
- Categorical Feature Selection (Chi-square)

#1.1

DATA EXPLORATION, CLEANSING, TRANSFORM

TOPICS

A. Data Exploration

B. Data Cleansing

C. Data Transform

LIBRARIES

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from sklearn import preprocessing

A. DATA EXPLORATION

1. Read .csv file
   - read_csv()
2. View Data Array Shape
   - # Variables
   - # Samples
   - Print()
3. View Variable info
   - Info()
     - Data Type
     - # non null

	X	Y	Z
0	19	1927	cat
1	NaN	2300	dog
2	15	NaN	bird
3	16	5959	cat
4	16	AB	cat
5	NaN	4594	dog
6	19	1927	cat
7	20	2879	bird��?
8	21	NaN	NaN
9	0	4096	cat
10	A	6730	cat
11	25	0	bird
12	0	2792	dog
13	33	2575	dog??
14	1000	4959	bird
15	19	1927	cat
16	36	4580	dog
17	40	5869	NaN
18	NaN	4178	dog
19	45	NaN	cat

(20, 3)

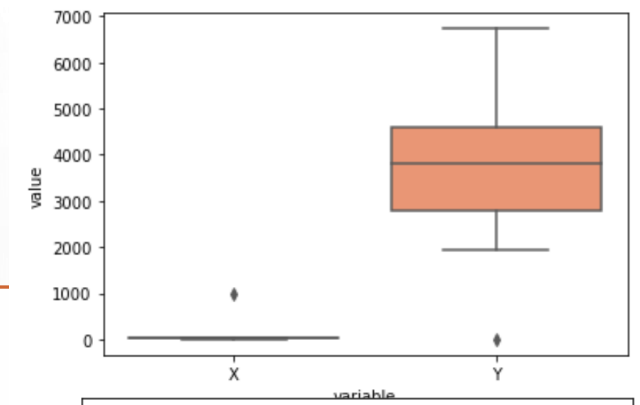
B. DATA CLEANSING

1. Correct Errors (delete non ASCII)
   - replace()
2. Convert data type from object to suitable types
   - X -> int64
   - Y -> float64
   - Z -> string
3. Drop duplicate samples (rows)
   - drop_duplicates()
4. View Variable Statistics
   - describe()
5. Drop rows with NA >1
   - dropna()
6. Replace NA
   - fillna()
     - X, Y with statistics mean or median
     - Z with previous rows

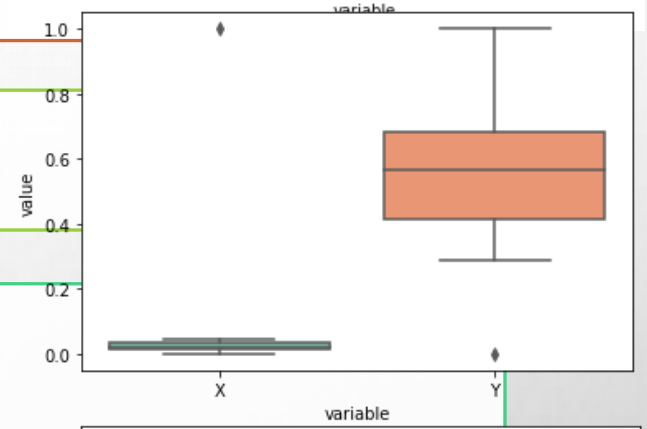
	X	Y	Z
0	19	1927.0	cat
1	20	2300.0	dog
2	15	3817.0	bird
3	16	5959.0	cat
4	16	3817.0	cat
5	20	4594.0	dog
7	20	2879.0	bird
9	0	4096.0	cat
10	20	6730.0	cat
11	25	0.0	bird
12	0	2792.0	dog
13	33	2575.0	dog
14	1000	4959.0	bird
16	36	4580.0	dog
17	40	5869.0	dog
18	20	4178.0	dog
19	45	3817.0	cat

C. DATA TRANSFORM

1
- Transform data
  - MinMaxScaler() / StandardScaler()

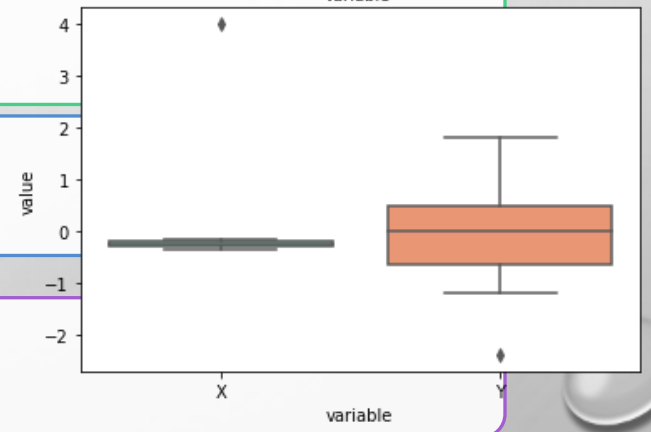


2
- Show Boxplot X,Y
  - sns.boxplot() or pd.boxplot()



3
- Remove outlier
  - Q1 = scaled_features[cols].quantile(0.25)
  - Q3 = scaled_features[cols].quantile(0.75)
  - IQR = Q3 - Q1
  - X, Y < Q1 – (1.5*IQR) | X,Y > Q3 + (1.5*IQR)

4
- Transform data
  - MinMaxScaler() / StandardScaler()



5
- Show Boxplot X,Y
- sns.boxplot() or pd.boxplot()

C. DATA CATEGORY LABEL

LabelEncoder() OneHotEncoder()

1. Reset drop index
   - reset_index()
2. Convert Category to Label
   - preprocessing.LabelEncoder()
3. Convert Category to Binary Label
   - preprocessing.OneHotEncoder()
4. Join LabelEncoder result, OneHotEncoder result to dataframe

	X	Y	Z	Z_category	bird	cat	dog
0	0.422222	0.000000	cat	1	0.0	1.0	0.0
1	0.444444	0.077660	dog	2	0.0	0.0	1.0
2	0.333333	0.393504	bird	0	1.0	0.0	0.0
3	0.355556	0.839475	cat	1	0.0	1.0	0.0
4	0.355556	0.393504	cat	1	0.0	1.0	0.0
5	0.444444	0.555278	dog	2	0.0	0.0	1.0
6	0.444444	0.198209	bird	0	1.0	0.0	0.0
7	0.000000	0.451593	cat	1	0.0	1.0	0.0
8	0.444444	1.000000	cat	1	0.0	1.0	0.0
9	0.000000	0.180096	dog	2	0.0	0.0	1.0
10	0.733333	0.134916	dog	2	0.0	0.0	1.0
11	0.800000	0.552363	dog	2	0.0	0.0	1.0
12	0.888889	0.820737	dog	2	0.0	0.0	1.0
13	0.444444	0.468665	dog	2	0.0	0.0	1.0
14	1.000000	0.393504	cat	1	0.0	1.0	0.0

#1.2

FEATURE SELECTION

TOPICS

A. Data Exploration, Cleansing, and Transform

B. Remove variables with High Variable Correlation

C. Variable Chi-square with High p-value

LIBRARIES

1. • import numpy as np
2. • import pandas as pd
3. • import matplotlib.pyplot as plt
4. • import seaborn as sns
5. • from sklearn import preprocessing
6. • from sklearn.feature_selection import chi2

A

Data Exploration, Cleansing, and Transform

A. Data exploration and Cleansing

1. Read .csv file
   - read_csv("Credit-Card-Defaulter-Prediction.csv",sep=",")
2. View Data Array Shape
   - # Variables
   - # Samples
3. Remove
   - 'customerID'
4. View Variable info
   - Info()
     - Data Type / # non null
5. Fill NA
   - fillna()

B.

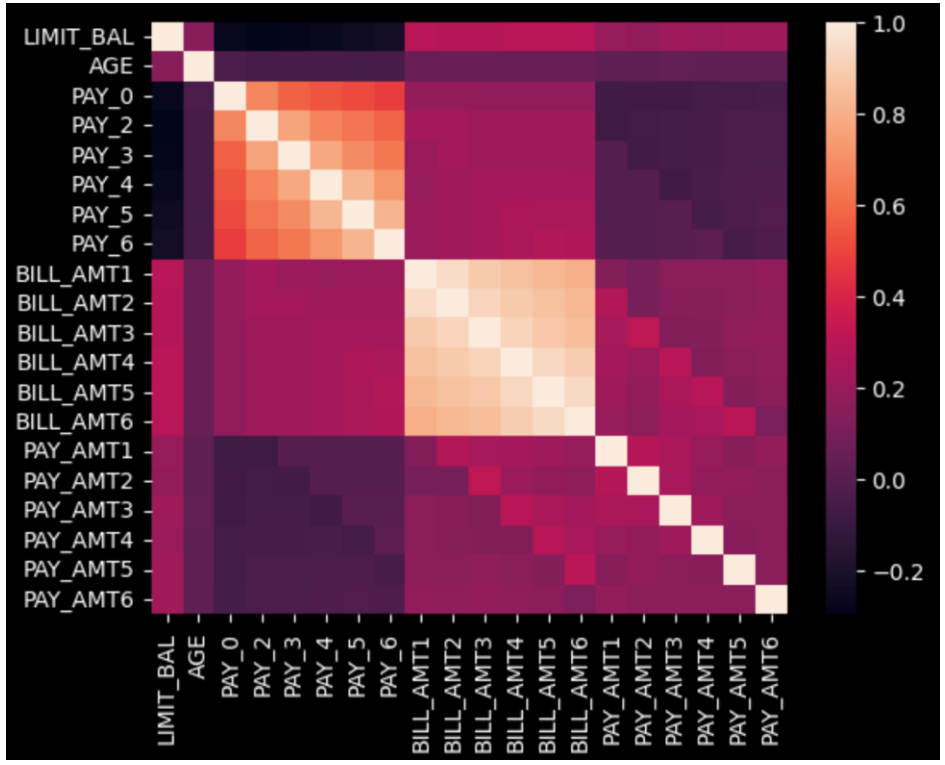
Remove variables with High Variable Correlation

B. Remove variables with High Variable Correlation

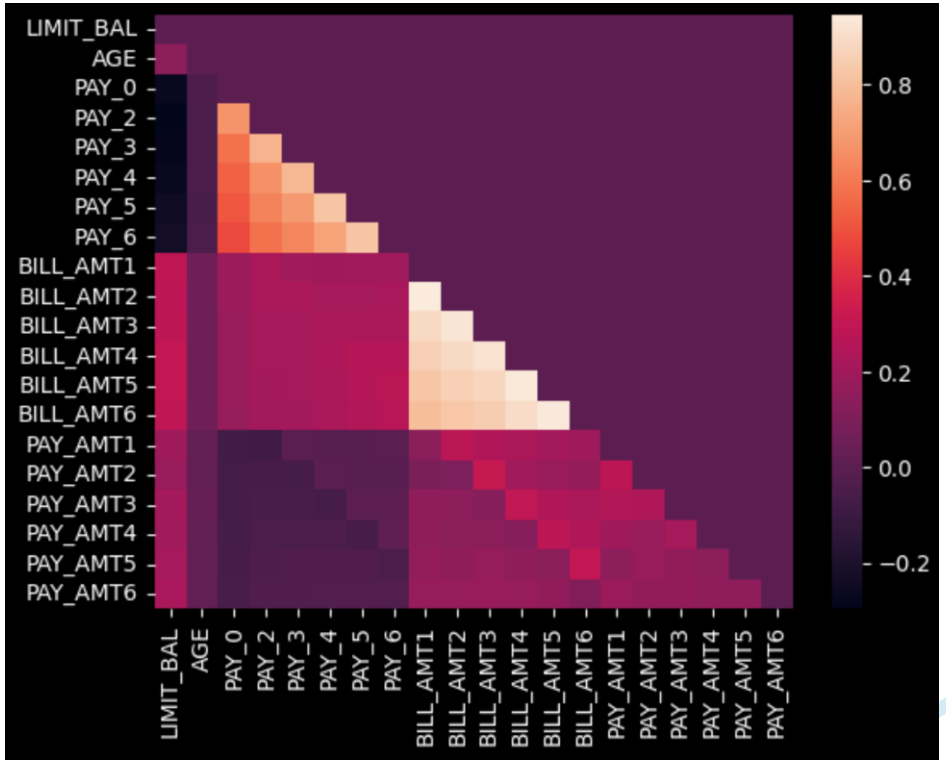
1. Create data frame of numerical data
2. Calculate correlation between variables (Corr())
3. Plot Headmap
   - Sns.heatmap()
4. Reduce Corr() to Lower Matrix (np.tril())
5. Drop columns if correlation value > 0.6
6. Show statistics of each numerical columns

B. Correlation (reduce to lower matrix)

BEFORE NP.TRIL()



AFTER NP.TRIL()



B. Results

	LIMIT_BAL	AGE	PAY_6	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
count	30000	30000	30000	30000	30000	3.00E+04	30000	30000	30000	30000
mean	167484.3227	35.4855	-0.2911	39693.41407	5817.551702	6.09E+03	5387.485652	4985.616598	4972.943356	5409.711199
std	129747.6616	9.217904	1.149988	59285.33997	16536.93438	2.30E+04	17582.93283	15641.56592	15251.0208	17748.95377
min	10000	21	-2	-339603	0	0.00E+00	0	0	0	0
25%	50000	28	-1	1760	1125	1.00E+03	614.75	416	412	389
50%	140000	34	0	18480.5	2400	2.20E+03	2000	1770.5	1880	1800
75%	240000	41	0	49198.25	5700	5.43E+03	5000	4985.616598	4972.943356	5000
max	1000000	79	8	961664	873552	1.68E+06	896040	621000	426529	528666

C.

Remove Variable with High p-value from Chi-square

C. Remove Variable with High p-value from Chi-square

1. Create data frame of Categorical data columns
2. Data Transform (Category to number) -> LabelEncoder()
3. Separate Input Variables and Output = [‘default’]
4. Show Chi-Square, p_value between input variables and output
5. Select insignificant variables with p_value > 0.05 (5%)
6. Create final data table [Numerical data, category data

C. Results

Drop p_value > 0.05

	LIMIT_BAL	SEX	EDUCATION	AGE	PAY_6	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default
0	20000	0	3	24	-2	39693.41407	5817.551702	689	0	0	0	0	1
1	120000	0	3	26	2	3261	0	1000	1000	1000	4972.943356	2000	1
2	90000	0	3	34	0	15549	1518	1500	1000	1000	1000	5000	0
3	50000	0	3	37	0	29547	2000	2019	1200	1100	1069	1000	0
4	50000	1	3	57	0	19131	2000	36681	10000	9000	689	679	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
29995	220000	1	1	39	0	15980	8500	20000	5003	3047	5000	1000	0
29996	150000	1	1	43	0	0	1837	3526	8998	129	0	5409.711199	0
29997	30000	1	3	37	0	19357	0	0	22000	4200	2000	3100	1
29998	80000	1	1	41	-1	48944	85900	3409	1178	1926	52964	1804	1
29999	50000	1	3	46	0	15313	2078	1800	1430	1000	1000	1000	1