Summary - Data Exploration, with Panda

conda env create = f environment.yml : create from f import pandse as pd import numpy as pp statiplotib inline import matplotib inline import matplotib inline import seatcer_matrix Source	ironment
Imported Packages Import pandas as pd import numpy as no handplottib inline import matplottib. pyplot as plt from pandas.plotting inport scatter_matrix **load_ext line_profiler import scatter_matrix** **def cat_ox (**application_train_cat_ext)** **load_ext line_profiler import scatter_matrix** **load_ext line_profiler import scatter_matrix** **load_ext line_profiler import imputes (**application_train_dat_ext_ext_ext_ext_ext_ext_ext_ext_ext_ex	
Read CSV df = pd.read_csv('application_train.csv')	iiie
Read CSV df = pd.read_csv('application_train.csv')	
Glimpse Data in Dataframe df.head(6)	
Dataframe Informtion df.info(null_counts=True, verbose=True) df.nunique().sort_values() df.nunique().sort_values() inter_cat = df.loc(;,df.nunique(axis=0).apply(lambod) Sort values inter_cat.nunique().sort_values() for c in inter_cat.columns:	
Distinct observations df.nunique().sort_values() variables that have 2 Values inter_cat = df.loc(:,df.nunique(axis=0).apply(lambod) for c in inter_cat.columns: print("(): \t {} ".format(c, df[c].unique())) cat_vars = \[\frac{1}{4}\].select_dtypes('object').columns.val *inter_cat.columns.values.tolist()] num_vars = \[\frac{1}{4}\].select_dtypes('object').columns.val *inter_cat.columns.values.tolist()] num_vars = \[\frac{1}{4}\].select_dtypes('object').columns.val *inter_cat.columns.difference(cat_vars).values.tolist()] num_vars = \[\frac{1}{4}\].select_olumns.difference(cat_vars).values.tolist()] num_vars = \[\frac{1}{4}\].select_olumns.difference(cat_vars).values.tolist()] volumns.values.tolist()] total = \[\frac{1}{4}\].sulue_counts() total = \[\frac{1}{4}\].sulue_().select_olumns.values.tolist()] percent = \[\langle \frac{1}{4}\].sulue_().select_olumns.values.tolist()] percent = \[\langle \frac{1}{4}\].sulue_().select_olumns.values.tolist()] total = \[\frac{1}{4}\].sulue_().select_olumns.values.tolist()] percent = \[\langle \frac{1}{4}\].sulue().select_olumns.values.tolist()] percent = \[\langle \frac{1}{4}\].sulue().select_olumns.values.tolist()] missing_application_train_data_query('Percent > 50') from_sklearn_preprocessing_import_Imputer_imputer_imputer = \[\langle \frac{1}{4}\].select_olumns.gelect_	
Variables that have 2 Values inter_cat = df.loc[:,df.nunique(axis=0).apply(lambod) Sort values Value of the Binary Variables Separate Categorical and Numeric Variables Class imbalance Overview of missing data Percentage of missing data Imputer Class Median Imputer Class Median Imputer Class Median Filling Missing Categorical Attributes with Special Value Option 1: pd.get_dummies (df[cat_vars], dummy_na=True): Option 2: encoder = LabelBinarizer() Encode Dummy Variables Option 2: encoder = LabelBinarizer() Data Distributions Descriptive Statistics Duild a Checklist for Exploratory Data Analysis Build a Checklist for Exploratory Data Analysis Separate Categorical Avars = Categorical Aralysis False) Inter_cat.columns.cript.org(, df.nunique(), ord (, square= True) inter_cat.columns.cript.org(, df.nunique(), ord (, square= True) inter_cat.columns.cript.org(, df.nunique(), ord (, square= True) imputer cat.columns.cript.org (, square= True) inter_cat.columns.cript.org (, square= True) imputer cat.columns.cript.org (, square= True) imputer cat.columns.cripts.cript.org (, square= True) imputer cat.columns.cripts.cript.org (, square= True) imputer cat.columns.cripts.cripts.cript.org (, square= True) imputer cat.columns.cripts.cri	
Sort values	
Value of the Binary Variables For c in inter_cat.columns:	da x: x < 3)]
Separate Categorical and Numeric Variables Separate Categorical and Numeric Variables Class imbalance Overview of missing data Percentage of missing data Fercentage of missing data Imputer Class Median Imputer Class Median Imputer Class Median Filling Missing Categorical Attributes with Special Value Option 1: pd.9et_dummy Variables Filling Missing Categorical Attributes with Special Value Option 1: pd.9et_dummies (df(cat_vars), dummy_na=True): Option 1: pd.9et_dummies (df(cat_vars), hist(figsize=(40,30)) plt. show() Data Distributions Descriptive Statistics Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extend of the power of the process in the power of the process of the category of the parallel of the power of the parallel of the parallel of the power of the parallel o	
Separate Categorical and Numeric Variables Num_vars = df.columns.difference(cat_vars).values.tolist()	
Class imbalance	
### Percentage of missing data Percentage of missing data	tolist()
Total = df.isnull().sum().sort_values(ascending = F percentage of missing data missing_application_train_data = pd.concat([total, keys=['Total', 'Percent']) missing_application_train_data.query('Percent > 50' from sklearn.preprocessing import Imputer imputer = Imputer(strategy="median", verbose= True) imputer.fit(df[num_vars]) X = imputer.transform(df[num_vars].copy()) temp = pd.DataFrame(X, columns=num_vars) Filling Missing Categorical Attributes with Special Value Option 1: pd.get_dummies(df[cat_vars], dummy_na=True) : Option 1: pd.get_dummies(df[cat_vars], dummy_na=True) : Option 2: encode Dummy Variables Option 2: encoder = LabelBinarizer() temp_cat_lhot = encoder.fit_transform(df['FLAG_DOCU df[num_vars].hist(figsize=(40,30)) plt.show() Data Distributions Descriptive Statistics df [num_vars].describe() Visualize Correlations Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extend	
False	False)
Imputer Class Median	, percent], axis=1,
Filling Missing Categorical Attributes with Special Value temp = df[cat_vars].fillna(999,inplace=False) Option 1: pd.get_dummies(df[cat_vars], dummy_na=True) : Option Option 2: encoder = LabelBinarizer() temp_cat_lhot = encoder.fit_transform(df['FLAG_DOCU df[num_vars].hist(figsize=(40,30)) plt.show() Descriptive Statistics df[num_vars].describe() Visualize Correlations plt.figure(figsize = (20,20)) ax = sns.heatmap(df[num_vars].corr(), square= True Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extendi	
### Descriptive Statistics Descriptive Statistics	
temp_cat_lhot = encoder.fit_transform(df['FLAG_DOCU df[num_vars].hist(figsize=(40,30)) plt.show() Descriptive Statistics df[num_vars].describe() Visualize Correlations plt.figure(figsize = (20,20)) ax = sns.heatmap(df[num_vars].corr(), square= True) Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extendi	on 1
Descriptive Statistics df [num_vars].describe() Visualize Correlations plt.figure(figsize = (20,20)) ax = sns.heatmap(df[num_vars].corr(), square= True) Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extendi	UMENT_4'])
df [num_vars].describe() Visualize Correlations plt.figure(figsize = (20,20)) ax = sns.heatmap(df[num_vars].corr(), square= True) Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extendi	
Feature Selection Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extendi	
Outliers Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extend	, linewidths=.5, cmap="Blues")
Build a Checklist for Exploratory Data Analysis Scaling the Analysis: How would you organize your analysis - extending the Analysis - extending t	
Scaling the Analysis: How would you organize your analysis - extend	
	ling it to include the other files in the data
Conversation Cafe Class Imbalance: In debt collection, data is often imbalance. How wo Missing Data: How do you decide on an appropriate Strategy to imp	•