In [133 In [134	<pre>import pandas as pd,numpy as np,os import matplotlib.pyplot as plt %matplotlib inline</pre>
In [134 In [135 Out[135	<pre>os.chdir('D:\machine learning\Raw data') os.listdir() ['a SQL', 'Advertising csy']</pre>
, J u	'Advertising.csv', 'automobile_data.sas7bdat', 'Automobile_data.xlsx', 'Automobile_data2.csv', 'Automobile_data2.xlsx', 'Automobile_data2.xlsx', 'bank_data.sas7bdat', 'bigmart_data.csv', 'Book1.xlsx',
	'carsnew2.xlsx', 'casnew.csv', 'churn.csv', 'churn.xlsx', 'churn2.csv', 'churn_data.pickle', 'churn_data.xlsx',
	<pre>'chur_12.xlsx', 'cleaned data', 'concrete_data.csv', 'Covid_data.xlsx', 'CREDIT_DISCOVERY_FOR_DS.csv', 'data.csv', 'data.csv', 'dubai_refreshments_final.sas7bdat', 'employees.csv',</pre>
	'employee_detail.sas7bdat', 'german.data.txt', 'Gold.xlsx', 'House Price.csv', 'House_Price_Scoring.csv', 'machine learning', 'MANJU.csv',
	<pre>'marks', 'merging', 'nortel.csv', 'payroll2.csv', 'Problem Statement.docx', 'state gdp', 'test.csv', 'Titanic_data.csv',</pre>
In [136	'train.csv', 'user_devise', 'user_usage.xlsx'] df=pd.read_csv('Titanic_data.csv',usecols=['Age','Fare','Survived'])
In [137 Out[137	<pre>df.head() Survived Age Fare 0 0 22.0 7.2500</pre>
	1 1 38.0 71.2833 2 1 26.0 7.9250 3 1 35.0 53.1000 4 0 35.0 8.0500
	<pre>df.isnull().sum() Survived 0 Age 177 Fare 0 dtype: int64</pre>
In [139 Out[139	df.isnull().mean() Survived 0.000000 Age 0.198653 Fare 0.000000
In [140 Out[140	<pre>dtype: float64 df['Age'].isnull().sum() 177</pre>
	177 50.00 305 0.92 292 36.00
	889 26.00 539 22.00 267 25.00 352 15.00 99 34.00 689 15.00 Name: Age, Length: 177, dtype: float64
In [142	<pre>df[df['Age'].isnull()].index Int64Index([5, 17, 19, 26, 28, 29, 31, 32, 36, 42,</pre>
In [143 Out[143	<pre>dtype='int64', length=177) df[df['Age'].isnull()].index Int64Index([5, 17, 19, 26, 28, 29, 31, 32, 36, 42,</pre>
In [144	<pre>dtype='int64', length=177) def impute_nan(df, variable, median): df[variable+"_median"]=df[variable].fillna(median) df[variable+"_random"]=df[variable] ##It will have the random sample to fill the na</pre>
In [145	<pre>random_sample=df[variable].dropna().sample(df[variable].isnull().sum(),random_state=0) ##pandas need to have same index in order to merge the dataset random_sample.index=df[df[variable].isnull()].index df.loc[df[variable].isnull(),variable+'_random']=random_sample median=df.Age.median()</pre> median=df.Age.median()
In [146 In [147	<pre>impute_nan(df, "Age", median) df.head()</pre>
Out[147	Survived Age Fare Age_median Age_random 0 0 22.0 7.2500 22.0 1 1 38.0 71.2833 38.0 2 1 26.0 7.9250 26.0
In [148	3 1 35.0 53.1000 35.0 35.0 4 0 35.0 8.0500 35.0 35.0 fig = plt.figure()
	<pre>ax = fig.add_subplot(111) df['Age'].plot(kind='kde', ax=ax) df.Age_median.plot(kind='kde', ax=ax, color='red') df.Age_random.plot(kind='kde', ax=ax, color='green') lines, labels = ax.get_legend_handles_labels() ax.legend(lines, labels, loc='best')</pre>
Out[148	0.05
	0.03 - 2 - 2 - 3 - 3 - 3 - 3 - 3 - 3 - 3 -
In [149	df=pd.read_csv('Titanic_data.csv', usecols=['Age', 'Fare', 'Survived']) df[:5]
Out[149	
	2 1 26.0 7.9250 3 1 35.0 53.1000 4 0 35.0 8.0500
In [150 In [151 Out[151	<pre>df['Age_new']=np.where(df['Age'].isnull(),1,0) df[:10] Survived Age Fare Age_new</pre>
	0 0 22.0 7.2500 0 1 1 38.0 71.2833 0 2 1 26.0 7.9250 0 3 1 35.0 53.1000 0
	4 0 35.0 8.0500 0 5 0 NaN 8.4583 1 6 0 54.0 51.8625 0 7 0 2.0 21.0750 0
T. 5450	8
Out[152 In [153 In [154	<pre>28.0 df["Age"].fillna(df["Age"].median(),inplace=True) df[:10]</pre>
Out[154	Survived Age Fare Age_new 0 0 22.0 7.2500 0 1 1 38.0 71.2833 0
	2 1 26.0 7.9250 0 3 1 35.0 53.1000 0 4 0 35.0 8.0500 0 5 0 28.0 8.4583 1 6 0 54.0 51.8625 0
	7 0 2.0 21.0750 0 8 1 27.0 11.1333 0 9 1 14.0 30.0708 0
	Advantages 1.Easy to implement 2.Captures the importance of missing values Disadvantages
In [155	1. Creating Additional Features(Curse of Dimensionality) End of Distribution imputation df=pd_read_csv('Titanic_data.csv', usecols=['Age', 'Fare', 'Survived'])
Out[155	Survived Age Fare
	1 1 38.0 71.2833 2 1 26.0 7.9250 3 1 35.0 53.1000 4 0 35.0 8.0500
In [156 Out[156	<pre>df.Age.hist(bins=50) <axessubplot:></axessubplot:></pre>
	40 30
In [157 In [158	<pre>extreme=df.Age.mean()+3*df.Age.std() import seaborn as sea sea.boxplot('Age', data=df)</pre>
	C:\Users\Admin\anaconda3\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional a rgument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation. warnings.warn(<axessubplot:xlabel='age'></axessubplot:xlabel='age'>
	0 10 20 30 40 50 60 70 80 Age
In [159	<pre>def impute_nan(df,variable, median, extreme): df['variable_extrm_val'] =df[variable].fillna(extreme) df[variable].fillna(median,inplace=True)</pre>
In [160 In [161 Out[161	<pre>impute_nan(df, 'Age', df.Age.median(), extreme) df.head(10) Survived Age Fare variable_extrm_val</pre>
	0 0 22.0 7.2500 22.00000 1 1 38.0 71.2833 38.00000 2 1 26.0 7.9250 26.00000 3 1 35.0 53.1000 35.00000
	4 0 35.0 8.0500 35.00000 5 0 28.0 8.4583 73.27861 6 0 54.0 51.8625 54.00000 7 0 2.0 21.0750 2.00000 8 1 27.0 11.1333 27.00000
In [162	9
Out[162	<pre><axessubplot:> 200 150</axessubplot:></pre>
	100
In [163	df['variable_extrm_val'].hist(bins=50) <axessuhplot:></axessuhplot:>
Out[163	175 150 125
	100 75 50 25
In [164	
	sea.boxplot('variable_extrm_val', data=df) C:\Users\Admin\anaconda3\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional a
	sea.boxplot('variable_extrm_val',data=df) C:\Users\Admin\anaconda3\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional a rgument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation. warnings.warn(
	sea.boxplot('variable_extrm_val',data=df) C:\Users\Admin\anaconda3\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional a rgument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation. warnings.warn(
	sea.boxplot('variable_extrm_val',data=df) C:\Users\ddis\name\name\name\name\name\name\name\name
	sea.boxplot('variable_extrm_val', data=df) C:\Users\Admin\anaconda3\lib\site-packages\Seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 8.12, the only valid positional a regument will be 'data', and passing other arguments without an explicit keyword will result in an error or misinterpretation. **AxesSubplot:xlabel='variable_extrm_val'> **AxesSubplot:xlabel='variable_extrm_val'> **Arbitrary Value Imputation this technique was derived from kaggle competition it consists of replacing NAN by an arbitrary value **df=pd.read.csv('Titanic_data.csv', usecols=['Age', 'Fare', 'Survived']) **df=pd.read.csv('Titanic_data.csv', usecols=['Age', 'Fare', 'Survived'])
In [165 Out[165	sea.boxplot('variable_extrm_val', data=df) C:\Users\Admin\anaconda3\):lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional a regument warnings.warn(dx.edu.leg.war.numbe.extrm_val Arbitrary Value Imputation this technique was derived from kaggle competition it consists of replacing NAN by an arbitrary value d=pd.read_tsv('Titanic_data.csv', usecols=['Age', 'Fare', 'Survived']) Survived Age Fare Survived Age Fare
In [165 Out[165	See a boxplot ("variable_exerm_val", dateoff) ChiterraVable in vancous and in Notice packages become in discretive any site bultumenaring. Place the full being variable as a Repeard arg: s. From version 9.12, the only valid positional a variable variable currently value in explicit keyword will result in an error or disinterpretation. Arbitrary Value Imputation this technique was derived from kaggle competition it consists of replacing NAN by an arbitrary value defind read_cay("fitanic data.csv", veccols=["Age", "Fare", "Survived"]) 5. Survived Age Fare 5. 0. 77.2 77.90 1. 1. 38.0 73.50 3. 1. 38.0 73.50 3. 1. 38.0 73.50 4. 0. 38.0 73.50 4. 0. 38.0 73.50 def impute new(df, variable): fill bag(s)
In [165 Out[165 In [166	Ditter's Normal mace and early Library decorations provided in the provided and an expectation of 22, the only valid positional and report of validable for a labble provided in the provided and the control of the provided in the provided
In [165 Out[165 In [166	College Andministration control value in sequence will be datal, and passing other arguments without an explicit support will be datal, and passing other arguments without an explicit support will be datal, and passing other arguments without an explicit support will be datal, and passing other arguments without an explicit support will be datal, and passing other arguments without an explicit support will be an under or mainterpretation. Arbitrary Value Imputation ### This technique was derived from knapple composition it explacing MAN by an arbitrary value #### This technique was derived from knapple composition it explacing MAN by an arbitrary value ###################################
In [165 Out[165 In [166	Southers American (a service services) - American (a services)
In [165 Out [165 In [167 Out [167 In [169	Sentence of the content of the conte
In [165 Out [165 In [167 Out [167 In [169	some heads of front into exercity and into exercity control of the
In [165 Out [165 In [167 Out [167 In [169	Substituting the (less after a cert cert) is transit. Children's certification of certification and certification in modificit expected will record to an accuracy or a hipport and certification. A certification of certification and certification
In [165 Out [165 In [167 Out [167 Out [169 Out [169	Section of the control
In [165 Out [165 In [167 Out [167 Out [169 Out [169	Exhance descriptions of the control
In [165 Out [165 In [167 Out [167 Out [169 Out [169	Exhance descriptions of the control
In [165 Out [165 In [167 Out [167 Out [169 Out [169	Exhance descriptions of the control
In [165 Out [165 In [167 Out [167 Out [169	Exhance descriptions of the control
In [165 Out [165 In [167 Out [167 Out [169 Out [169	Exhance descriptions of the control