

Data Mining

Lab - 5 - Data Preprocessing

Jeet Bhalodi (23031701006)

1) First, you need to read the titanic dataset from local disk and display Last five records

```
In [2]: import pandas as pd
In [4]: df = pd.read_csv('titanic.csv')
df
```

Harris Cumings, Mrs. John Bradley (Florence Briggs Th 2] :	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
1	0	1	0	3	Mr. Owen	male	22.0	1	0		7.2!
2 3 1 3 Miss. Italian 26.0 0 0 3 SION/OZ. 7.3101282 7.3	1	2	1	1	Mrs. John Bradley (Florence Briggs	female	38.0	1	0	PC 17599	71.2{
Mrs. Jacques Heath (Lily May Peel) Female 35.0 1 0 113803 53.2	2	3	1	3	Miss.	female	26.0	0	0		7.9%
4 5 0 3 William Henry male 35.0 0 0 373450 8.5	3	4	1	1	Mrs. Jacques Heath (Lily May	female	35.0	1	0	113803	53.1(
886 887 0 2 Rev. Juozas male 27.0 0 0 211536 13. 887 888 1 1 Graham, Miss. Margaret Edith female 19.0 0 0 112053 30. 888 889 0 3 Catherine Helen "Carrie" female NaN 1 2 W./C. 6607 23. 889 890 1 1 Karl Howell Howell male 26.0 0 0 111369 30. 890 891 0 3 Mr. Patrick male 32.0 0 0 370376 7.	4	5	0	3	William	male	35.0	0	0	373450	8.0!
886 887 0 2 Rev. male 27.0 0 0 211536 13. 887 888 1 1 Amagaret Edith Female 19.0 0 0 112053 30. 888 889 0 3 Catherine Helen "Carrie" NaN 1 2 W./C. 6607 23. 889 890 1 1 Karl Marl Male 26.0 0 0 111369 30. 890 891 0 3 Mr. Mr. Male 32.0 0 0 370376 7. Patrick 891 0 3 Mr. Mr. Male 32.0 0 0 370376 7.	•••	•••	•••		•••						
887 888 1 1 Miss. Margaret Edith female 19.0 0 0 112053 30.0 888 889 0 3 Catherine Helen "Carrie" female NaN 1 2 W./C. 6607 23.0 889 890 1 1 Karl Howell male 26.0 0 0 111369 30.0 890 891 0 3 Mr. Patrick male 32.0 0 0 370376 7.0	886	887	0	2	Rev.	male	27.0	0	0	211536	13.00
888 889 0 3 Catherine Helen "Carrie" female NaN 1 2 W./C. 6607 23. 889 890 1 1 Behr, Mr. Karl Howell male 26.0 0 0 111369 30. 890 891 0 3 Dooley, Patrick male 32.0 0 0 370376 7.0 Patrick 7 Patrick 9 0 0 370376 7.0	887	888	1	1	Miss. Margaret	female	19.0	0	0	112053	30.00
889 890 1 1 1 Karl male 26.0 0 0 111369 30. Howell Dooley, 890 891 0 3 Mr. male 32.0 0 0 370376 7. Patrick	888	889	0	3	Miss. Catherine Helen	female	NaN	1	2		23.4!
890 891 0 3 Mr. male 32.0 0 0 370376 7. Patrick	889	890	1	1	Karl	male	26.0	0	0	111369	30.00
891 rows × 12 columns	890	891	0	3	Mr.	male	32.0	0	0	370376	7.7!
4	891	rows × 12 colu	mns								
	4										•

In [6]: df.tail(5)

Out[6]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare (
	886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
	887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45
	889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
	890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75
	4										•

2) Handle Missing Values in data set [use dropna(), fillna(), and interpolate]

In [12]: dropna = df.dropna()
dropna

L2]:		Passengerld	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.283
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100
	6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.862
	10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.700
	11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.550
	•••										
8	871	872	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47.0	1	1	11751	52.554
8	872	873	0	1	Carlsson, Mr. Frans Olof	male	33.0	0	0	695	5.000
8	879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.158
8	887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000
8	889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000
18	83 rc	ows × 12 colur	nns								

•

```
In [19]: dropna_how = df.dropna(how='any',axis=1)
dropna_how
```

Out[19]:		Passengerld	Survived	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare
	0	1	0	3	Braund, Mr. Owen Harris	male	1	0	A/5 21171	7.2500
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	1	0	PC 17599	71.2833
	2	3	1	3	Heikkinen, Miss. Laina	female	0	0	STON/O2. 3101282	7.9250
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	1	0	113803	53.1000
	4	5	0	3	Allen, Mr. William Henry	male	0	0	373450	8.0500
	•••		•••	•••		•••			•••	
	886	887	0	2	Montvila, Rev. Juozas	male	0	0	211536	13.0000
	887	888	1	1	Graham, Miss. Margaret Edith	female	0	0	112053	30.0000
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	1	2	W./C. 6607	23.4500
	889	890	1	1	Behr, Mr. Karl Howell	male	0	0	111369	30.0000
	890	891	0	3	Dooley, Mr. Patrick	male	0	0	370376	7.7500

891 rows × 9 columns

```
In [23]: # if row has all missing value so it is remove
dropna_how_all = df.dropna(how='all',axis=0)
```

dropna_how_all

3]: _		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
	0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2!
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.28
	2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9%
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1(
	4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0!
	•••										
	886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
i	887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
i	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4!
;	889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
	890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7!
8	91 rc	ows × 12 colur	nns								
	→										

```
In [31]: df_filna = df.fillna('xyz')
    df_filna
```

]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fi
	0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.28
	2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
	4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	20.8
	•••										
	886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
	887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	xyz	1	2	W./C. 6607	23.45
	889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
	890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75
8	891 rc	ows × 12 colur	nns								

```
df_filna.info()
In [33]:
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 891 entries, 0 to 890
       Data columns (total 12 columns):
            Column
                         Non-Null Count Dtype
            -----
                         -----
                                         ____
        0
            PassengerId 891 non-null
                                         int64
        1
            Survived
                         891 non-null
                                         int64
            Pclass
                         891 non-null
                                         int64
         3
            Name
                         891 non-null
                                         object
        4
            Sex
                         891 non-null
                                         object
         5
                         891 non-null
                                         object
            Age
                                         int64
         6
            SibSp
                         891 non-null
         7
            Parch
                         891 non-null
                                         int64
            Ticket
                         891 non-null
                                         object
        9
            Fare
                         891 non-null
                                         float64
        10 Cabin
                         891 non-null
                                         object
        11 Embarked
                         891 non-null
                                         object
       dtypes: float64(1), int64(5), object(6)
       memory usage: 83.7+ KB
In [55]: data_fillna1 = df.fillna({'Age': 18, 'Cabin': 'Not Available'})
         data_fillna1
```

Harris Cumings, Mrs. John Bradley Female 38.0 1 0 PC 17599 71.2 Bernham Production of the production o	•	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fi
1 2 1 1 Mrs. John Bradley (Florence Briggs Th female Bradley (Florence Briggs Th 3 4 1 3 Heikkinen, Mrs. Laina female Bradley (Florence Briggs Th 6 2 0 0 STON/O2 3101282 7.9 3 4 1 1 Heikkinen, Mrs. Mrs. Laina female Bradley (Florence Briggs Th 6 2 5 0 3 Allen, Mr. William Henry Peel) 7.9 7.9 7.2 4 5 0 3 Allen, Mr. William Henry Peel) 7.0 0 0 373450 8.0 886 887 0 2 Rev. Juozas Miss. Margaret Edith male 27.0 0 0 211536 13.0 887 888 1 1 Miss. Miss. Margaret Edith 6 1 2 W./C. 6607 23.4 888 889 0 3 3 6 1 1 2 W./C. 6607 23.4 890 891 0 3 Behr, Mr. Patrick <th< td=""><td>0</td><td>1</td><td>0</td><td>3</td><td>Mr. Owen</td><td>male</td><td>22.0</td><td>1</td><td>0</td><td></td><td>7.25</td></th<>	0	1	0	3	Mr. Owen	male	22.0	1	0		7.25
2 3 1 3 Miss. female 26.0 0 0 3 101/02.2 7.9 Futrelle, Mrs. Jacques Heath (Lily May Peel) 4 5 0 3 William male 35.0 0 0 373450 8.0 Montvila, Miss. Margaret Edith 887 888 1 1 1 Miss. Margaret Edith 888 889 0 3 Catherine Helen "Carrie" 889 890 1 1 Behr, Mr. Mr. Karl Howell 890 891 0 3 Mr. male 26.0 0 0 370376 7.7 Booley, Bell Miss. Margaret Melen 35.0 0 0 370376 7.7 Booley, Margaret Market Market Market Market Melen Miss. Margaret Miss. Margaret Melen Miss. Margaret Miss. Marga	1	2	1	1	Mrs. John Bradley (Florence Briggs	female	38.0	1	0	PC 17599	71.28
Mrs. Jacques Heath (Lily May Peel) Female 35.0 1 0 113803 53.1	2	3	1	3	Miss.	female	26.0	0	0		7.92
4 5 0 3 William Henry male 35.0 0 0 373450 8.0	3	4	1	1	Mrs. Jacques Heath (Lily May	female	35.0	1	0	113803	53.10
886 887 0 2 Montvila, Rev. Juozas male 27.0 0 0 211536 13.0 887 888 1 1 Graham, Miss. Margaret Edith female 19.0 0 0 112053 30.0 888 889 0 3 Catherine Helen "Carrie" female 18.0 1 2 W/C. 6607 23.4 889 890 1 1 Karl Howell male 26.0 0 0 111369 30.0 890 891 0 3 Mr.	4	5	0	3	William	male	35.0	0	0	373450	20.8
886 887 0 2 Rev. male 27.0 0 0 211536 13.0 887 888 1 1 Graham, Miss. Margaret Edith female 19.0 0 0 112053 30.0 888 889 0 3 Catherine Helen "Carrie" female 18.0 1 2 W./C. 6607 23.4 889 890 1 1 Karl Howell male 26.0 0 0 111369 30.0 890 891 0 3 Mr. Mr. Mr. Mr. Mr. Mr. Male 32.0 0 0 370376 7.7 891 rows × 12 columns	•••										
887 888 1 1 Miss. Margaret Edith female 19.0 0 0 112053 30.0 888 889 0 3 Catherine Helen "Carrie" female 18.0 1 2 W./C. 6607 23.4 889 890 1 1 Karl Mr. Mark Mr. Male 26.0 0 0 111369 30.0 890 891 0 3 Mr. Mr. Male 32.0 0 0 370376 7.7 891 rows × 12 columns	886	887	0	2	Rev.	male	27.0	0	0	211536	13.00
888 889 0 3 Catherine Helen "Carrie" female 18.0 1 2 W./C. 6607 23.4 889 890 1 1 Behr, Mr. Mr. Male 26.0 0 0 111369 30.0 890 891 0 3 Mr. Mr. Male 32.0 0 0 370376 7.7 891 rows × 12 columns	887	888	1	1	Miss. Margaret	female	19.0	0	0	112053	30.00
889 890 1 1 1 Karl male 26.0 0 0 111369 30.0 Howell Dooley, Patrick 891 rows × 12 columns	888	889	0	3	Miss. Catherine Helen	female	18.0	1	2		23.45
890 891 0 3 Mr. male 32.0 0 0 370376 7.7 Patrick 891 rows × 12 columns	889	890	1	1	Karl	male	26.0	0	0	111369	30.00
	890	891	0	3	Mr.	male	32.0	0	0	370376	7.75
	891 rd	ows × 12 colur	mns								
	4										•

```
In [49]: meanAge = df.Age.mean()
meanAge

Out[49]: 29.69911764705882

In [67]: data_fillna2 = df.fillna({'Age': meanAge, 'Cabin': 'Not Available'})
data_fillna2
```

67]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
	0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.000000	1	0	PC 17599
	2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803
	4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450
	•••									
	886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536
	887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607
	889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369
	890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376
8	891 rd	ows × 12 colur	mns							
	4									•

```
In [77]: cabbinMod = df.Cabin.mode()[0][0:3]
Out[77]: 'B96'
In [79]: data_fillna3 = df.fillna({'Age': meanAge, 'Cabin':cabbinMod })
    data_fillna3
```

79]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
	0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.000000	1	0	PC 17599
	2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803
	4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450
	•••		•••		•••	•••				
	886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536
	887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607
	889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369
	890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376
8	391 rd	ows × 12 colur	mns							
	4									•

```
In [87]: data_interpolate = df.interpolate()
    data_interpolate
```

C:\Users\ASUS\AppData\Local\Temp\ipykernel_15020\2280711911.py:1: FutureWarning: Dat
aFrame.interpolate with object dtype is deprecated and will raise in a future versio
n. Call obj.infer_objects(copy=False) before interpolating instead.
 data_interpolate = df.interpolate()

7]:		PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fi
	0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
	1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.28
	2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
	3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
	4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05
	•••										
	886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00
	887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00
	888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	22.5	1	2	W./C. 6607	23.45
	889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00
	890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75
8	891 rc	ows × 12 colur	mns								
	√										•

3) Apply Scaling to AGE attribute with min max, decimal scaling and z score.

```
# MIN-MAX Normalization
     In [94]:
                                                    df['AGE_minmax'] = (df['Age'] - df['Age'].min()) / (df['Age'].max() - df['Age'].min()) / (df['Age'].min()) / (df['Age'].min(
In [109...
                                                  df[['Age','AGE_minmax']]
Out[109...
                                                                              Age AGE_minmax
                                                               0 22.0
                                                                                                                           0.271174
                                                                              38.0
                                                                                                                           0.472229
                                                                              26.0
                                                                                                                           0.321438
                                                               3 35.0
                                                                                                                           0.434531
                                                               4 35.0
                                                                                                                           0.434531
                                                    886
                                                                              27.0
                                                                                                                           0.334004
                                                    887 19.0
                                                                                                                           0.233476
                                                    888 NaN
                                                                                                                                             NaN
                                                    889 26.0
                                                                                                                           0.321438
                                                    890 32.0
                                                                                                                           0.396833
                                                891 rows × 2 columns
In [119...
                                                    # Decimal Normalization
                                                    import numpy as np
                                                    temp = len(str(int(df['Age'].max())))
                                                    df['AGE_decimal'] = df['Age'] / (10**temp)
                                                 df[['Age','AGE_minmax']]
In [129...
```

Out[129		Age	AGE_minmax
	0	22.0	0.271174
	1	38.0	0.472229
	2	26.0	0.321438
	3	35.0	0.434531
	4	35.0	0.434531
	•••		
	886	27.0	0.334004
	887	19.0	0.233476
	888	NaN	NaN
	889	26.0	0.321438
	890	32.0	0.396833
	891 rc	ows × i	2 columns
In [98]:			Normalizatio
In [98]:			
	df['/	AGE_zs	score'] = (df
In [131	df['/	AGE_zs 'Age',	score'] = (di
In [131	df['/	AGE_zs 'Age',	core'] = (df
In [131	df['/	AGE_zs 'Age', Age	'AGE_zscore' AGE_zscore
In [131	df[['/	AGE_zs 'Age', Age 22.0	AGE_zscore
In [131	df[['/ df[[0 1 2	AGE_zs 'Age', Age 22.0 38.0	'AGE_zscore' AGE_zscore -0.530005 0.571430
In [131	df[['/ df[[0 1 2	AGE_zs 'Age', Age 22.0 38.0 26.0	'AGE_zscore' AGE_zscore -0.530005 0.571430 -0.254646
In [131	df['/ df[[0 1 2 3	AGE_zs 'Age', Age 22.0 38.0 26.0 35.0	AGE_zscore -0.530005 0.571430 -0.254646 0.364911
In [131	df['/ df[[0 1 2 3 4	AGE_zs 'Age', Age 22.0 38.0 26.0 35.0	'AGE_zscore' AGE_zscore -0.530005 0.571430 -0.254646 0.364911 0.364911
In [98]: In [131 Out[131	df['/ df[[0 1 2 3 4	AGE_zs 'Age', Age 22.0 38.0 26.0 35.0	AGE_zscore -0.530005 0.571430 -0.254646 0.364911 0.364911
In [131	df['/ df[[0 1 2 3 4 886	AGE_zs 'Age', Age 22.0 38.0 26.0 35.0 27.0	AGE_zscore -0.530005 0.571430 -0.254646 0.364911 0.364911 -0.185807

891 rows × 2 columns

0.158392

890 32.0

In []:	
In []:	
In []:	