Exploratory Data Analysis (EDA)

Written by: M.Danish Azeem Date: 08.12.2023 Email: danishazeem365@gmail.com

```
# import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1- import dataset

```
df = sns.load dataset("titanic")
df
     survived pclass
                                  age sibsp parch
                                                          fare embarked
                            sex
class
                          male
                                 22.0
                                                        7.2500
                                                                       S
Third
                        female
                                 38.0
                                                       71.2833
                                                                       C
1
             1
First
                        female
                                 26.0
                                                        7.9250
                                                                       S
Third
                        female
                                 35.0
                                                       53.1000
                                                                       S
First
                                                                       S
                          male
                                 35.0
                                                   0
                                                        8.0500
Third
. .
886
                          male
                                 27.0
                                                       13.0000
                                                                       S
Second
                        female
                                                       30.0000
                                                                       S
                                 19.0
887
First
                        female
                                                      23.4500
                                                                       S
888
                     3
                                  NaN
Third
                          male
                                 26.0
                                                      30.0000
                                                                       C
889
First
890
                          male 32.0
                                                        7.7500
Third
             adult male deck
                               embark_town alive
       who
                                                   alone
0
                   True
                         NaN
                               Southampton
                                                   False
       man
                                               no
1
                  False
                            C
                                 Cherbourg
                                              yes
                                                   False
     woman
2
                  False
                         NaN
                               Southampton
                                                    True
     woman
                                              yes
3
                  False
                            C
                               Southampton
                                                   False
                                              yes
     woman
```

```
4
                     True
                            NaN
                                  Southampton
                                                         True
                                                   no
        man
        . . .
                                                   . . .
                                                          . . .
                      . . .
886
                     True
                            NaN
                                  Southampton
                                                         True
        man
                                                   no
887
                    False
                              В
                                  Southampton
     woman
                                                  yes
                                                         True
888
                    False
                            NaN
                                  Southampton
                                                        False
     woman
                                                   no
889
                              C
                                                         True
        man
                     True
                                    Cherbourg
                                                  yes
890
        man
                     True
                            NaN
                                   Queenstown
                                                   no
                                                         True
[891 rows \times 15 columns]
```

1- Bigger picture of data

```
df.dtypes
survived
                   int64
pclass
                   int64
sex
                  object
                 float64
age
sibsp
                   int64
                   int64
parch
fare
                 float64
embarked
                  object
class
                category
who
                  object
adult male
                    bool
deck
                category
embark town
                  object
alive
                  object
alone
                    bool
dtype: object
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
#
     Column
                   Non-Null Count
                                    Dtype
 0
     survived
                   891 non-null
                                    int64
                   891 non-null
 1
     pclass
                                    int64
 2
                   891 non-null
                                    object
     sex
 3
                   714 non-null
                                    float64
     age
 4
                   891 non-null
                                    int64
     sibsp
 5
                   891 non-null
     parch
                                    int64
 6
     fare
                   891 non-null
                                    float64
 7
                                    object
     embarked
                   889 non-null
 8
     class
                   891 non-null
                                    category
 9
     who
                   891 non-null
                                    object
```

```
10
     adult male
                  891 non-null
                                   bool
                  203 non-null
 11
     deck
                                   category
 12
     embark town
                  889 non-null
                                   object
 13
     alive
                  891 non-null
                                   object
 14
     alone
                  891 non-null
                                   bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
df.isnull().sum()
                    # finding missing and null values
survived
                 0
                 0
pclass
                 0
sex
               177
age
                 0
sibsp
parch
                 0
                 0
fare
                 2
embarked
                 0
class
                 0
who
adult male
                 0
               688
deck
embark town
                 2
                 0
alive
                 0
alone
dtype: int64
df2 = df
df2
     survived pclass
                          sex
                                 age sibsp parch fare embarked
class \
            0
                    3 male 22.0
                                          1
                                                  0
                                                      7.2500
                                                                    S
0
Third
                                                                    \mathsf{C}
                      female 38.0
                                          1
                                                    71.2833
First
                       female
                                26.0
                                                                    S
            1
                                                      7.9250
Third
                    1 female
                                35.0
                                          1
                                                     53.1000
                                                                    S
3
            1
                                                  0
First
                         male 35.0
                                                      8.0500
                                                                    S
Third
. . .
886
            0
                          male 27.0
                                          0
                                                     13.0000
                                                                    S
Second
                       female 19.0
                                                                    S
887
                                          0
                                                     30.0000
First
            0
                                          1
                                                                    S
888
                    3 female
                                 NaN
                                                  2 23.4500
```

```
Third
                           male
                                  26.0
                                             0
                                                        30.0000
                                                                        C
889
             1
                      1
First
                           male 32.0
890
             0
                      3
                                             0
                                                         7.7500
                                                                        0
Third
             adult male deck
                                embark_town alive
                                                     alone
       who
0
       man
                   True
                          NaN
                                Southampton
                                                no
                                                     False
1
                  False
                                  Cherbourg
     woman
                            C
                                               yes
                                                     False
2
                  False
                          NaN
                                Southampton
                                                     True
     woman
                                               yes
3
                  False
                            C
                                Southampton
                                                     False
                                               yes
     woman
4
                                Southampton
                                                     True
                   True
       man
                          NaN
                                                no
                     . . .
                                               . . .
                                                       . . .
        . . .
                          . . .
886
                          NaN
                                Southampton
                                                      True
                   True
       man
                                                no
887
     woman
                   False
                            В
                                Southampton
                                               yes
                                                     True
888
                          NaN
                                Southampton
                                                     False
     woman
                  False
                                                no
                   True
                                  Cherboura
                                                      True
889
                            C
       man
                                               ves
890
                   True
                          NaN
                                 Queenstown
                                                      True
       man
                                                no
[891 rows x 15 columns]
df2.isnull().sum() / len(df2) * 100
                 0.000000
survived
                 0.000000
pclass
                 0.000000
sex
                19.865320
age
sibsp
                 0.000000
                 0.000000
parch
                 0.000000
fare
embarked
                 0.224467
class
                 0.000000
                 0.000000
who
adult male
                 0.000000
                77.216611
deck
embark town
                 0.224467
alive
                 0.000000
alone
                 0.000000
dtype: float64
```

#1 Assighnment : How to deal with missing values in EDA Analysis. like categorical/object, numeric(int,float) , boolean etc

```
print(df2["age"].max())
print(df2["age"].min())
print(df2["age"].mean())

80.0
0.42
29.69911764705882
```

```
df2['age'].unique()
                    , 26.
                            , 35.
array([22.
             , 38.
                                        nan, 54.
                                                     2.
                                                          , 27.
                                                                  , 14.
                    , 20.
                            , 39.
                                           , 31.
             , 58.
                                                            15.
                                                                  , 28.
        4.
                                   , 55.
                                                    34.
                                   , 42.
        8.
             , 19.
                    , 40.
                            , 66.
                                             21.
                                                     18.
                                                             3.
                                                                     7.
                            , 28.5
       49.
             , 29.
                                       5.
                                             11.
                                                     45.
                      65.
                                                            17.
                                                                    32.
       16.
               25.
                       0.83, 30.
                                     33.
                                             23.
                                                     24.
                                                            46.
                                                                    59.
       71.
               37.
                            , 14.5 , 70.5
                                             32.5
                                                    12.
                                                             9.
                      47.
                                                                    36.5
                                           , 61.
       51.
               55.5
                      40.5 , 44.
                                     1.
                                                     56.
                                                            50.
                                                                    36.
       45.5 ,
               20.5
                                    , 52.
                                             63.
                                                    23.5
                      62.
                             41.
                                                             0.92, 43.
             , 10.
                    , 64.
                            , 13.
                                   , 48.
                                              0.75, 53.
                                                            57.
                                                                  , 80.
       60.
                               0.67, 30.5, 0.42, 34.5, 74.
       70.
             , 24.5
                    , 6. ,
df2['age'].fillna(df['age'].mean(), inplace=True)
df2
     survived pclass
                                                               fare
                            sex
                                        age
                                             sibsp
                                                    parch
embarked \
                     3
                           male
                                 22.000000
                                                             7.2500
S
1
                        female
                                 38.000000
                                                            71.2833
C
2
                                 26.000000
                     3
                        female
                                                             7.9250
S
3
                        female
                                 35.000000
                                                            53.1000
S
4
                           male
                                 35.000000
                                                             8.0500
S
. .
886
                     2
                           male
                                 27,000000
                                                            13.0000
S
887
                        female
                                 19.000000
                                                            30.0000
                                                 0
S
888
                        female
                                 29.699118
                                                            23.4500
S
889
                           male
                                 26,000000
                                                            30,0000
C
890
                     3
                           male 32.000000
                                                             7.7500
0
      class
                     adult male deck
                                        embark town alive
                                                            alone
                who
      Third
                            True
                                        Southampton
                                                            False
0
                                  NaN
                man
                                                        no
1
      First
                           False
                                    C
                                          Cherbourg
                                                            False
              woman
                                                       yes
2
      Third
                           False NaN
                                        Southampton
                                                             True
              woman
                                                       yes
3
      First
              woman
                           False
                                    C
                                        Southampton
                                                       yes
                                                            False
4
      Third
                            True
                                  NaN
                                        Southampton
                                                             True
                man
                                                       no
                . . .
                                                       . . .
                                                               . . .
886
                            True
                                  NaN
                                        Southampton
     Second
                                                       no
                                                             True
                man
```

887 Fir									
888 Thi	rd wo	oman oman	False False		Soutl	nampton nampton	yes no	True False	
889 Fir		man	True	С		erbourg	yes	True	
890 Thi	.rd	man	True	NaN	Que	enstown	no	True	
[891 rows	x 15	columns							
df2.isnul	l(). <mark>su</mark>	ım()							
survived pclass sex age sibsp parch fare embarked class who adult_mal deck embark_to alive alone dtype: in	own	0 0 0 0 0 0 2 0 0 0 688 2 0							
		pclass	sex		age	sibsp	parch	fare	
embarked	\		_						
0	^		7						
	0	3	male	22.00	0000	1	0	7.2500	
S									
S 1	1	3 1	female	38.00		1	0	7.2500 71.2833	
S 1	1	1	female	38.00	0000	1	0	71.2833	
S 1 C 2					0000				
S 1 C 2	1	1	female	38.00 26.00	0000 0000	1 0	0 0	71.2833	
S 1 C 2	1	1	female	38.00	0000 0000	1	0	71.2833	
S 1 C 2	1 1 1	1 3 1	female female	38.00 26.00 35.00	0000 0000 0000	1 0 1	9 9 9	71.2833 7.9250 53.1000	
S 1 C 2	1	1	female	38.00 26.00	0000 0000 0000	1 0	0 0	71.2833	
S 1 C 2 S 3 S 4 S	1 1 1	1 3 1	female female	38.00 26.00 35.00	0000 0000 0000	1 0 1	9 9 9	71.2833 7.9250 53.1000	
S 1 C 2	1 1 1	1 3 1	female female	38.00 26.00 35.00	0000 0000 0000	1 0 1	9 9 9	71.2833 7.9250 53.1000	
S 1 C 2 S 3 S 4 S	1 1 1	1 3 1	female female	38.00 26.00 35.00	0000 0000 0000 0000	1 0 1	9 9 9	71.2833 7.9250 53.1000	
S 1 C 2 S 3 S 4 S 886 S	1 1 1 0	1 3 1 3	female female male male	38.00 26.00 35.00 35.00	0000 0000 0000 0000	1 0 1 0	9 9 0 0	71.2833 7.9250 53.1000 8.0500	
S 1 C 2 S 3 S 4 S 886 S 887	1 1 1 0	1 3 1 3	female female male	38.00 26.00 35.00 35.00	0000 0000 0000 	1 0 1 0	9 9 0 0	71.2833 7.9250 53.1000 8.0500	
S 1 C 2 S 3 S 4 S 886 S 887 S	1 1 0 	1 3 1 3 2	female female male male female	38.00 26.00 35.00 35.00 27.00	0000 0000 0000 0000	1 0 1 0 0	9 9 0 9	71.2833 7.9250 53.1000 8.0500 13.0000 30.0000	
S 1 C 2 S 3 S 4 S 886 S 887 S 888	1 1 1 0 	1 3 1 3 	female female male male	38.00 26.00 35.00 35.00	0000 0000 0000 0000	1 0 1 0	0000	71.2833 7.9250 53.1000 8.0500 13.0000	
S 1 C 2 S 3 S 4 S 886 S 887 S	1 1 0 	1 3 1 3 2	female female male male female	38.00 26.00 35.00 35.00 27.00	0000 0000 0000 0000 0000 9118	1 0 1 0 0	9 9 0 9	71.2833 7.9250 53.1000 8.0500 13.0000 30.0000	

890		0	3	male	32.00	0000 0	0	7.7500
Q								
	class	who	adul	t_male		embark_town		alone
0	Third	man		True	NaN	Southampton	no	False
1	First	woman		False	C	Cherbourg	yes	False
2	Third	woman		False	NaN	Southampton	yes	True
3	First	woman		False	C	Southampton	yes	False
4	Third	man		True	NaN	Southampton	no	True
886	Second	man		True	NaN	Southampton	no	True
887	First	woman		False	В	Southampton	yes	True
888	Third	woman		False	NaN	Southampton	no	False
889	First	man		True	C	Cherbourg	yes	True
890	Third	man		True	NaN	Queenstown	no	True
_		_						
[891	rows x	15 colu	mns]					
df3	= df2							
a i J	u i Z							

df3

survived pclass sex age sibsp parch embarked \ 0 0 3 male 22.000000 1 0	fare 7.2500 71.2833	
0 0 3 male 22.000000 1 0		
S 1 1 formula 20 000000 1 0	71.2833	
1 1 f_{omp} 20 000000 1 0	71.2833	
1 1 female 38.000000 1 0		
	7.9250	
2 1 3 female 26.000000 0 0 S S 1 1 female 35.000000 1 0 S	53.1000	
4 0 3 male 35.000000 0 0 S	8.0500	
886 0 2 male 27.000000 0 0	13.0000	
S 887 1 1 female 19.000000 0 0	30.0000	
S 888 0 3 female 29.699118 1 2	23.4500	
S 889 1 1 male 26.000000 0 0	30.0000	
C 890 0 3 male 32.000000 0 0	7.7500	
Q		
class who adult_male deck embark_town alive Third man True NaN Southampton no First woman False C Cherbourg yes		

```
2
      Third
                           False
                                   NaN
                                        Southampton
                                                               True
                                                        ves
              woman
3
      First
                                     C
                                        Southampton
                                                              False
                           False
              woman
                                                        yes
4
      Third
                            True
                                   NaN
                                        Southampton
                                                         no
                                                               True
                man
                . . .
                              . . .
                                                        . . .
                                                                . . .
886
     Second
                            True
                                   NaN
                                        Southampton
                                                         no
                                                               True
                man
887
      First
                           False
                                     В
                                        Southampton
                                                              True
              woman
                                                        yes
                                                              False
888
      Third
                           False
                                  NaN
                                        Southampton
              woman
                                                         no
889
      First
                            True
                                     C
                                           Cherbourg
                                                               True
                                                        yes
                man
890
      Third
                            True NaN
                                          Queenstown
                man
                                                         no
                                                               True
[891 rows x 15 columns]
df3.isnull().sum()
survived
                  0
                  0
pclass
                  0
sex
                  0
age
sibsp
                  0
                  0
parch
                  0
fare
embarked
                  2
                  0
class
                  0
who
adult male
                  0
                688
deck
embark_town
                  2
                  0
alive
                  0
alone
dtype: int64
df3 = df3.drop(columns=["deck"])
df3.isnull().sum()
survived
                0
pclass
                0
                0
sex
age
                0
                0
sibsp
                0
parch
fare
                0
embarked
                2
class
                0
                0
who
                0
adult male
                2
embark town
                0
alive
alone
                0
dtype: int64
```

```
df4 = df3
df4["embarked"].fillna(df["embarked"].mode()[0], inplace=True)
df4["embark town"].fillna(df["embark town"].mode()[0], inplace=True)
df4.isnull().sum()
survived
               0
               0
pclass
               0
sex
               0
age
               0
sibsp
               0
parch
fare
               0
embarked
               0
class
who
adult male
embark town
               0
alive
               0
alone
               0
dtype: int64
print("mean = ",df['age'].mean())
print("median = ",df['age'].median())
print("mode = ",df['embarked'].mode())
mean = 29.69911764705882
median = 29.69911764705882
mode = 0
Name: embarked, dtype: object
```

steps: Data wranglings (EDA)

- 1. Import labireses
- 2. Import database
- 3. Explore your data
 - a. information
 - b. Datatype
 - c. Missing values
 - d. Take sence of your data
- 4. Understanding the variables
- 5. Relationship between the variables Analysts (heatmap, pairplot, correlation)
- 6. Brainstorming
 - a. Normalize (Technics # asignment)
 - b. Removing outliers # Asighnment
- 7. Tidy data, clean data

- 8. Ready for statstitical Analystis
- 9. Ready for Predection
- 10. Ready for machin learning
- 11. Ready for DL.