Assignment 2 Draft saved

File Edit View Run Add-ons Help

Share | Save Version 0

Run All | Markdown | Draft Session (28m)

[1]:
```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/titanic/train.csv
/kaggle/input/titanic/test.csv
/kaggle/input/titanic/gender_submission.csv
/kaggle/input/testdata/test.csv
```

Reading the CSV file in concern

+ Code + Markdown

[2]:
```
df = pd.read_csv('/kaggle/input/titanic/train.csv')
```

[3]:
```
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

[4]:
```
df.shape
```

[4]: (891, 12)

Checking the null values in each column

[5]:
```
total_null = df.isnull().sum().sort_values(ascending = False)
percent = ((df.isnull().sum()/df.isnull().count())*100).sort_values(ascending = False)
print("Total records = ", df.shape[0])

missing_data = pd.concat([total_null,percent.round(2)],axis=1,keys=['Total Missing','In Percent'])
missing_data.head(10)
```

```
Total records =  891
```

	Total Missing	In Percent
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22
PassengerId	0	0.00
Survived	0	0.00
Pclass	0	0.00
Name	0	0.00
Sex	0	0.00
SibSp	0	0.00
Parch	0	0.00

Fixing the NULL values in the columns

[6]:
```
df = df.drop(['Cabin'], axis = 1)
```

[7]:
```
df['Age'].fillna(value = df['Age'].mean().round(0), inplace=True)
```

[13]:
```
df = df.dropna(subset=['Embarked'])
```

[14]:
```
total_null = df.isnull().sum().sort_values(ascending = False)
percent = ((df.isnull().sum()/df.isnull().count())*100).sort_values(ascending = False)
print("Total records = ", df.shape[0])

missing_data = pd.concat([total_null,percent.round(2)],axis=1,keys=['Total Missing','In Percent'])
missing_data.head(10)
```

Total records = 889

[14]:

	Total Missing	In Percent
PassengerId	0	0.0
Survived	0	0.0
Pclass	0	0.0
Name	0	0.0
Sex	0	0.0
Age	0	0.0
SibSp	0	0.0
Parch	0	0.0
Ticket	0	0.0
Fare	0	0.0

[8]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

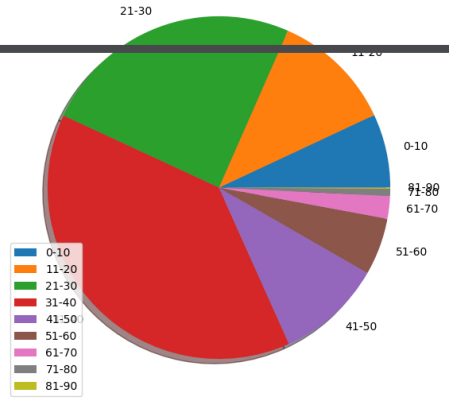
[9]:
```
survived_count = df.groupby('Survived')['Survived'].count() # can be done using value counts
survived_count
```

[9]:
```
Survived
0    549
1    342
Name: Survived, dtype: int64
```

[ ]:

Pie chart for different age groups

[10]:
```
 agedist=df[df['Age'].notnull()]['Age'].values
ages=np.histogram(agedist,bins=[0,10,20,30,40,50,60,70,80,90])
ages_label=['0-10','11-20','21-30','31-40','41-50','51-60','61-70','71-80','81-90']
plt.figure(figsize=(7, 7))
explode=[0,0,0,0,0,0,0,0,0]
plt.pie(x=ages[0], labels=ages_label,shadow=True,explode=explode)
plt.legend()
plt.show()
```



[ ]: