## Future Intern Project of Data Analytics Task 2

# Task 1: Calculate summary statistics (Mean, Median, Mode, Standard Deviation For a dataset

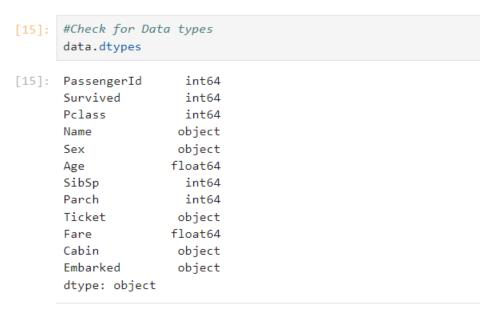
# Steps:

## 1. Import Packages and Train dataset and load and display

#### 1. Train Dataset



## 2. First rows of Dataset



## 3. Get Identify Categorical and Numerical Columns

```
•[17]: #Idetify the Categorical Columns
       categorical_columns = data.select_dtypes (include=['object','category']).columns.tolist()
       #Identify the Numerical Columns
       numerical_columns = data.select_dtypes(include=['int64','float64']).columns.tolist()
[23]: print("Categorical_columns:")
       print(categorical_columns)
       Categorical_columns:
       ['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked']
[25]: print("Numerical_columns")
       print(numerical_columns)
       Numerical_columns
       ['PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
[35]: columns_used = ['PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
       selected_columns = data.loc[:, columns_used]
[37]: selected_columns.head(5)
          PassengerId Survived Pclass Age SibSp Parch
                                                            Fare
       0
                   1
                             0
                                                         7.2500
                                    3 22.0
                                       38.0
                                                       0 71.2833
                                                          7.9250
       2
                   3
                             1
                                    3 26.0
                                                0
                                    1 35.0
                                                       0 53.1000
       3
       4
                   5
                             0
                                    3 35.0
                                                0
                                                       0 8.0500
```

#### 4. Lets check for mean for each column

```
#Lets check for mean for each column
column_means = selected_columns.mean()
print("Mean for each column:",column_means)
Mean for each column: PassengerId 446.000000
Survived
                0.383838
Pclass
                2.308642
               29.699118
Age
SibSp
                0.523008
Parch
                0.381594
               32.204208
Fare
dtype: float64
```

#### 5. Lets check for mode for each column

```
#Lets check for mode for each column
column_modes = selected_columns.mode().iloc[0]
print("Mode for each column:",column_modes)
Mode for each column: PassengerId
                                     1.00
Survived
              0.00
Pclass
              3.00
Age
             24.00
SibSp
               0.00
Parch
               0.00
               8.05
Name: 0, dtype: float64
```

## 6. Lets check for median for each column

```
#Lets check for median for each column
column median = selected columns.median()
print("Median for each column:",column_median)
Median for each column: PassengerId
                                       446,0000
                0.0000
Survived
Pclass
                 3.0000
Age
                28.0000
                0.0000
SibSp
Parch
                 0.0000
Fare
                14.4542
dtype: float64
```

#### 7. Lets check for standard deviation for each column

```
#Lets check for standard deviation for each column
column_std_deviation = selected_columns.std()
print("Standard Deviation for each column:",column_std_deviation)
Standard Deviation for each column: PassengerId
                                                  257.353842
Survived
               0.486592
Pclass
                0.836071
               14.526497
Age
SibSp
                1.102743
Parch
                0.806057
Fare
               49.693429
dtype: float64
```

# 8. Check for statistics for each numeric column all together

#Check for statistics for each nnumeric column all together selected\_columns.describe() SibSp PassengerId Survived Pclass Age Parch Fare count 891.000000 891.000000 891.000000 714.000000 891.000000 891.000000 mean 446.000000 0.383838 2.308642 29.699118 0.523008 0.381594 32.204208 0.486592 std 257.353842 0.836071 14.526497 1.102743 0.806057 49.693429 min 1.000000 0.000000 1.000000 0.420000 0.000000 0.000000 0.000000 25% 223.500000 0.000000 2.000000 20.125000 0.000000 0.000000 7.910400 446.000000 0.000000 3.000000 28.000000 0.000000 0.000000 14.454200 50% **75**% 668.500000 1.000000 3.000000 38.000000 1.000000 0.000000 31.000000 891.000000 1.000000 3.000000 80.000000 8.000000 6.000000 512.329200

#### 9. Check for Stats data



```
•[53]: #Check for two unique variables
       data['Sex'].unique()
[53]: array(['male', 'female'], dtype=object)
•[55]: #We have two variables
       data['Survived'].unique()
[55]: array([0, 1], dtype=int64)
[59]: data[["Sex","Age"]].groupby("Sex").mean()
[59]:
                   Age
          Sex
       female 27.915709
         male 30.726645
[61]: data[["Survived", "Age"]].groupby("Survived").mean()
[61]:
                     Age
       Survived
             0 30.626179
             1 28.343690
 [63]: data.groupby(["Sex","Survived"])["Age"].mean()
              Survived
 [63]: Sex
        female 0
                          25.046875
                          28.847716
               1
                          31.618056
        male
              1
                          27.276022
       Name: Age, dtype: float64
```

## By

**Shaikh Jasmin Kauser**