Big Data Tools for Managers

Dept. of MBA, Siddaganga Institute of Technology-Tumkur

Working with Titanic Dataset

The files we just opened are available on the data page for the Titanic competition on Kaggle. That page also has a data dictionary, which explains the various columns that make up the data set. Below are the descriptions contained in that data dictionary:

Download Dataset:

Link-1:
https://raw.githubusercontent.com/sitmbadept/sitmbadept.github.io/main/BDTM/R/titanic.csv

Link-2: https://drive.google.com/file/d/1LsrhPCyKceXWhtlqdGaYkSO_ufH6G8da/view?usp=sharing

Note: This link take you in another page and then click on File menu -> Save Page As -> Open file window and click on Save

- **PassengerID**— A column added by Kaggle to identify each row and make sublessons easier
- **Survived**— Whether the passenger survived or not and the value we are predicting (0=No, 1=Yes)
- **Pclass**— The class of the ticket the passenger purchased (1=1st, 2=2nd, 3=3rd)
- **Name**- The name of passenger's
- **Sex**— The passenger's sex
- **Age**— The passenger's age in years
- **SibSp**— The number of siblings or spouses the passenger had aboard the Titanic
- **Parch**— The number of parents or children the passenger had aboard the Titanic
- **Ticket**— The passenger's ticket number
- **Fare**— The fare the passenger paid
- **Cabin**— The passenger's cabin number
- **Embarked**— The port where the passenger embarked (C=Cherbourg, Q=Queenstown, S=Southampton)

Write Python Code for below questions.

1. Import pandas library in Python
2. Read titanic dataset in Python
3. Get the dimension of Titanic dataset
4. Display column names of dataset
5. View data
6. Get Quick summary for all the columns
7. Indentify the Null(Missing) values for dataset
8. How many male and female are on Titanic?
9. Find out maximum Ticket Fare
10. How many passengers got survived according to dataset
11. Count the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)

Note : This analysis is based on very few variables, you may have to practice yourself with remaing variables to get more hands-on with R & Data

In [ ]:
```
# Import pandas library in Python

# !pip install pandas #Execute this statement if you don't have download & install pand
import pandas as pd
```

In [ ]:
```
# Read titanic dataset in Python
data = pd.read_csv("https://raw.githubusercontent.com/sitmbadept/sitmbadept.github.io/m
```

In [ ]:
```
# Get the dimension of Titanic dataset
data.shape
```

In [ ]:
```
# Display column names of dataset
data.columns
```

In [ ]:
```
# View data
print(data)
```

In [ ]:
```
# Get Quick summary for all the columns
data.describe(include='all')
```

In [ ]:
```
# Indentify the Null(Missing) values for dataset
data.isnull().sum()
```

In [ ]:
```
# How many male and female are on Titanic?
data['Sex'].value_counts()
```

In [ ]:
```
# Find out maximum Ticket Fare
data['Fare'].max()
```

In [ ]:
```
#How many passengers got survived according to dataset
data['Survived'].value_counts()
```

In [ ]:
```
# Count the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)
data['Pclass'].value_counts()
```