

DR. ALVIN'S PUBLICATIONS

DATAPREP AI AND MISSING NO

WITH PYTHON
DR. ALVIN ANG



1 | PAGE

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

CONTENTS

I. Pip Install	3
II. Import Libraries	4
A. Import DataPrep	4
B. Import MissingNo	4
C. Load Dataset.....	4
III. MSNO	5
IV. DataPrep - Overview	7
V. DataPrep – Missing Values	8
VI. DataPrep - Correlation	10
VII. DataPrep – Describing a Single (Numerical) Column	12
VIII. DataPrep – Describing a Single (Categorical) Column	13
IX. DataPrep – Describing Two Numerical Columns	14
X. DataPrep – Describing Two Columns (Numerical vs Categorical)	15
XI. DataPrep – Describing Two Categorical Columns	16
XII. DataPrep – Compare Two Dataframes	17
XIII. DataPrep – Create HTML Report	18
XIV. DataPrep – Analyze Geographical Data	19
About Dr. Alvin Ang	20

I. PIP INSTALL

https://www.alvinang.sg/s/DataPrep_MissingNo_by_Dr_Alvin_Ang.ipynb

References:

<https://towardsdatascience.com/using-the-missingno-python-library-to-identify-and-visualise-missing-data-prior-to-machine-learning-34c8c5b5f009#:~:text=Missingno%20is%20an%20excellent%20and,%2C%20heatmap%2C%20or%20a%20dendrogram>

<https://dataprep.ai/>

1) Pip Install

```
[ ] pip install -U dataprep
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: dataprep in /usr/local/lib/python3.7/dist-packages (0.4.3)
Requirement already satisfied: python-stddnum<2.0,>=1.16 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.17)
Requirement already satisfied: nltk<4.0.0,>=3.6.7 in /usr/local/lib/python3.7/dist-packages (from dataprep) (3.7)
Requirement already satisfied: wordcloud<2.0,>=1.8 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.8.2.2)
Requirement already satisfied: flask_cors<4.0.0,>=3.0.10 in /usr/local/lib/python3.7/dist-packages (from dataprep) (3.0.10)
Requirement already satisfied: numpy<2.0,>=1.21 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.21.6)
Requirement already satisfied: ipywidgets<8.0,>=7.5 in /usr/local/lib/python3.7/dist-packages (from dataprep) (7.7.1)
Requirement already satisfied: pydantic<2.0,>=1.6 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.9.2)
Requirement already satisfied: tqdm<5.0,>=4.48 in /usr/local/lib/python3.7/dist-packages (from dataprep) (4.64.1)
Requirement already satisfied: python-Levenshtein<0.13.0,>=0.12.2 in /usr/local/lib/python3.7/dist-packages (from dataprep) (0.12.2)
Requirement already satisfied: python-cvtsuite<0.10.0,>=0.9.7 in /usr/local/lib/python3.7/dist-packages (from dataprep) (0.9.8)
Requirement already satisfied: jsonpath-ng<2.0,>=1.5 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.5.3)
Requirement already satisfied: scipy<=1.7.1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.7.1)
Requirement already satisfied: regex<2022.0.0,>=2021.8.3 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2021.11.10)
Requirement already satisfied: aiohttp<4.0,>=3.6 in /usr/local/lib/python3.7/dist-packages (from dataprep) (3.8.1)
Requirement already satisfied: flask<3,>=2 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2.2.2)
Requirement already satisfied: pandas<2.0,>=1.1 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.3.5)
Requirement already satisfied: dask[array,dataframe,delayed]<2022.0,>=2021.11 in /usr/local/lib/python3.7/dist-packages (from dataprep) (2021.12.0)
Requirement already satisfied: sqlalchemy<2.0.0,>=1.4.32 in /usr/local/lib/python3.7/dist-packages (from dataprep) (1.4.41)
Requirement already satisfied: metaphone<0.7,>=0.6 in /usr/local/lib/python3.7/dist-packages (from dataprep) (0.6)
```

II. IMPORT LIBRARIES

A. IMPORT DATAPREP

2) Import Libraries

2a) Import DataPrep

```
[ ] from dataprep.datasets import load_dataset
    from dataprep.eda import plot, plot_correlation, plot_missing, plot_diff, create_report
```

B. IMPORT MISSINGNO

2b) Import MissingNo

```
[ ] import pandas as pd
    import missingno as msno
    import matplotlib.pyplot as plt
    %matplotlib inline
```

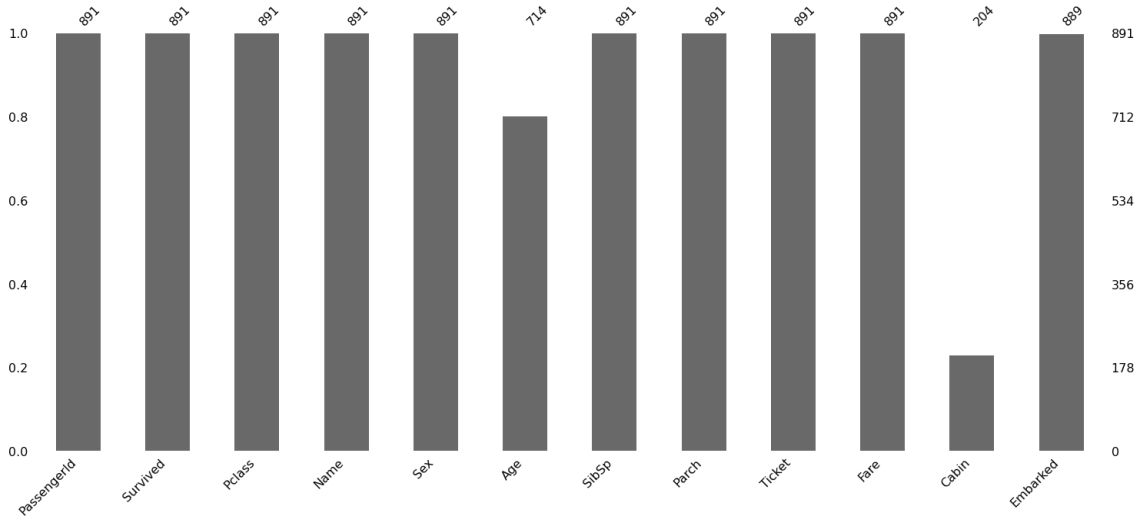
C. LOAD DATASET

2c) Load Dataset

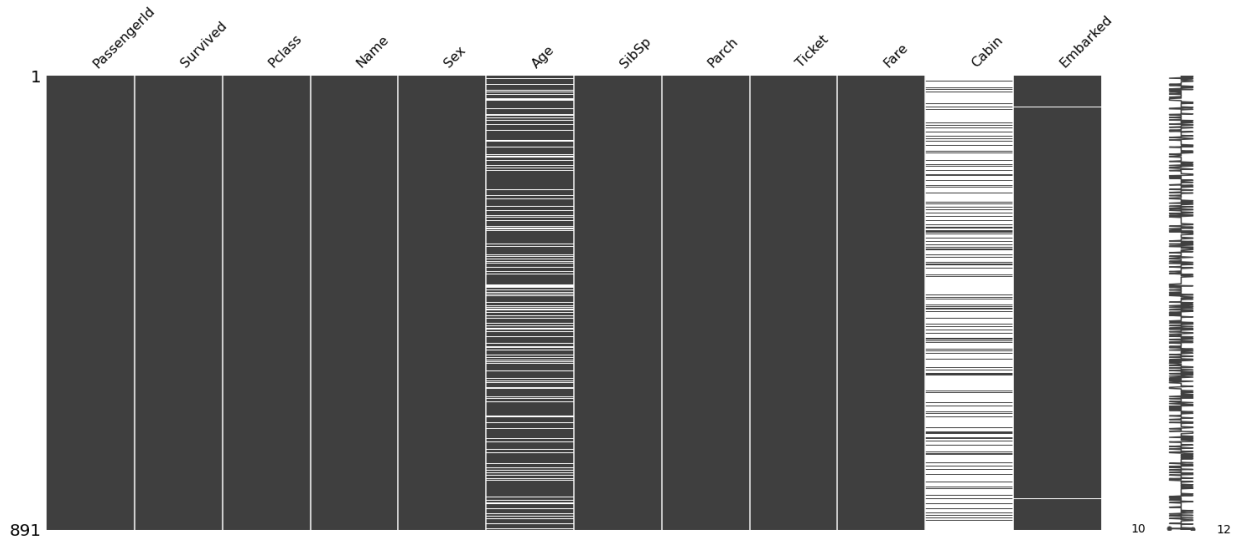
```
[ ] df = load_dataset("titanic")
```

III. MSNO

```
3) MSNO  
▶ msno.bar(df)
```



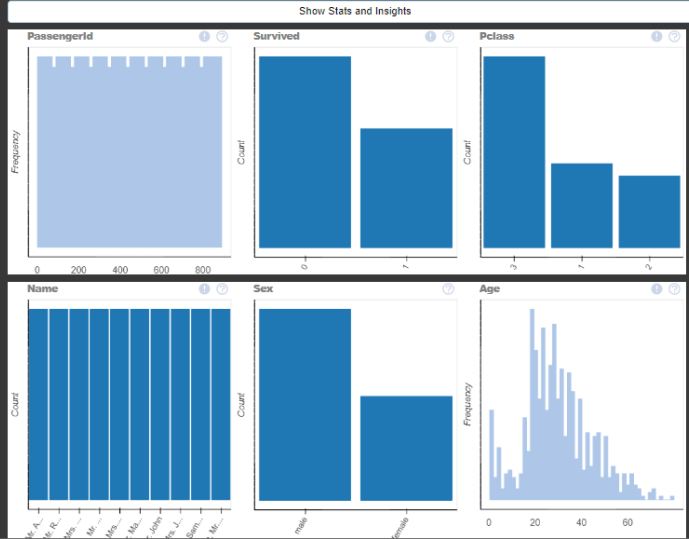
```
msno.matrix(df)
```



IV. DATAPREP - OVERVIEW

4) DataPrep - Overview

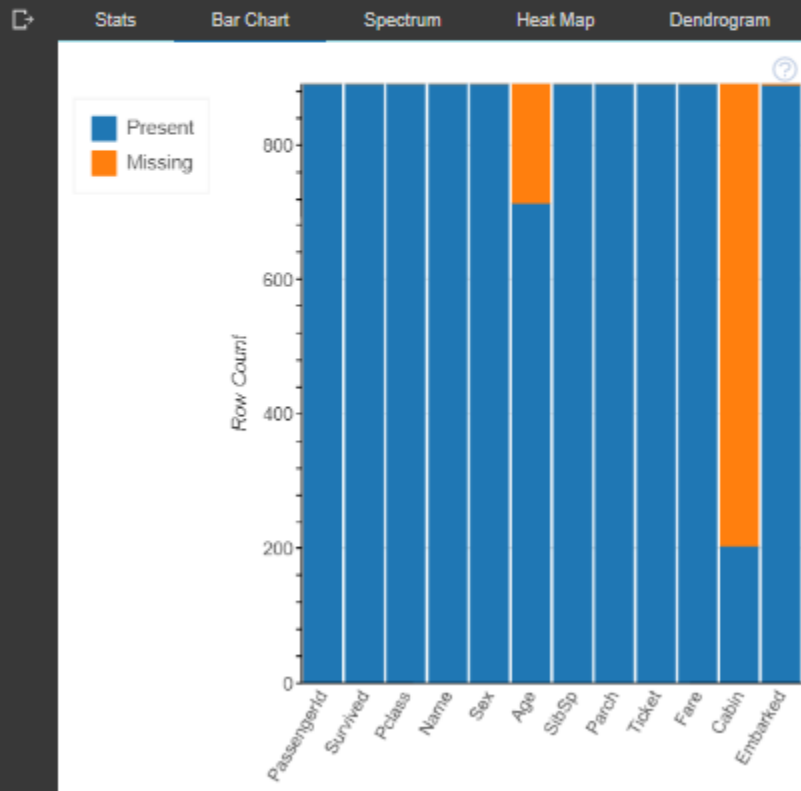
plot(df)



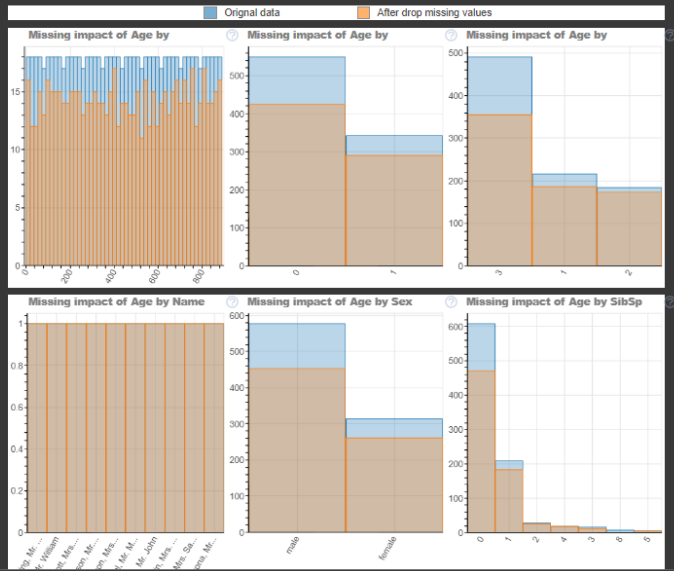
5) DataPrep - Missing Values

```
plot_missing(df)
```

```
#missing rows overview
```



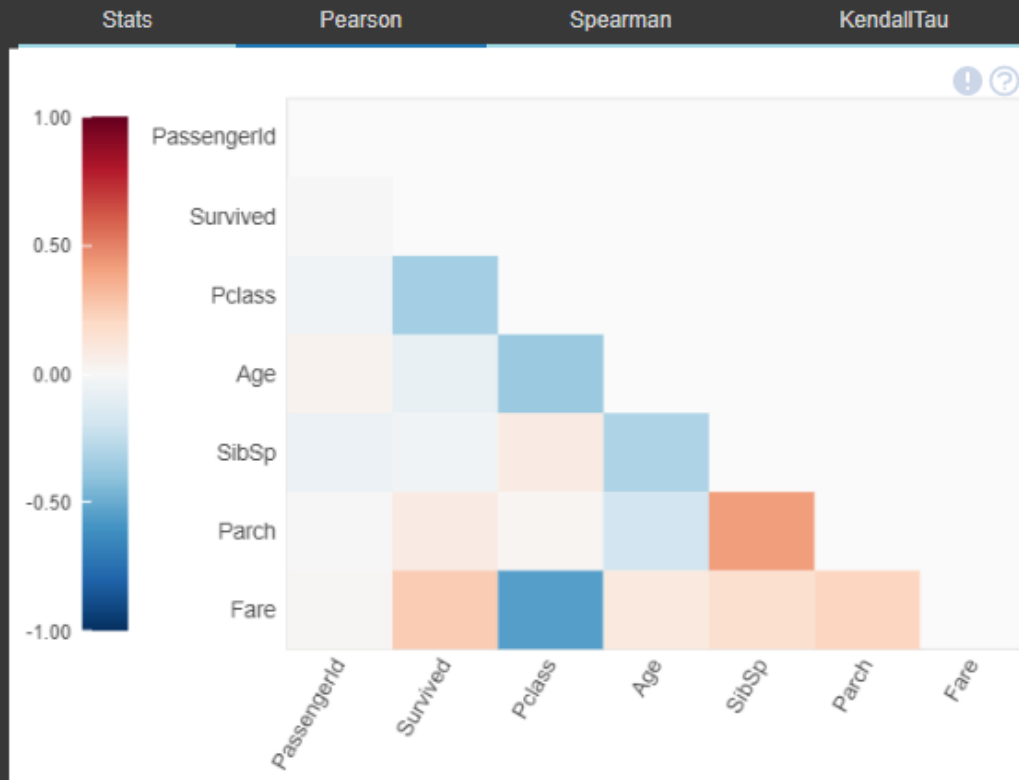

```
[ ] plot_missing(df, 'Age ')  
  
#understand how the other columns change  
#after dropping the missing values of 'age' column
```



6) DataPrep - Correlation

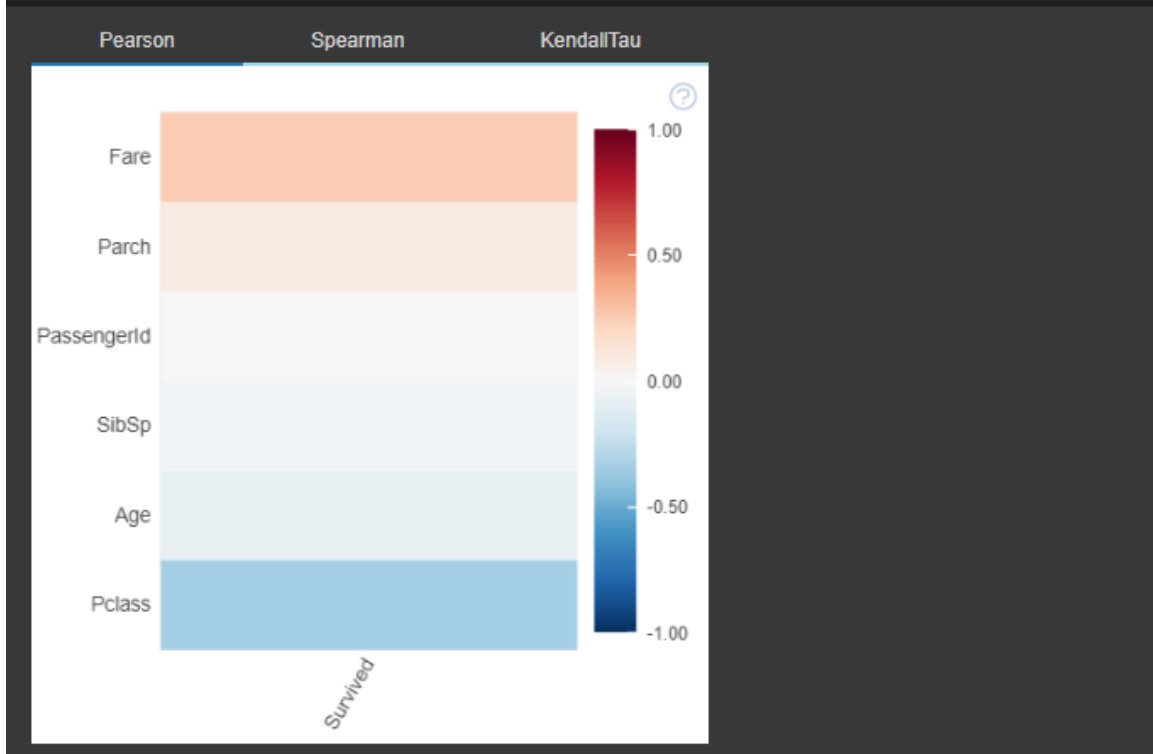
```
▶ plot_correlation(df)
```

```
#correlation overview
```



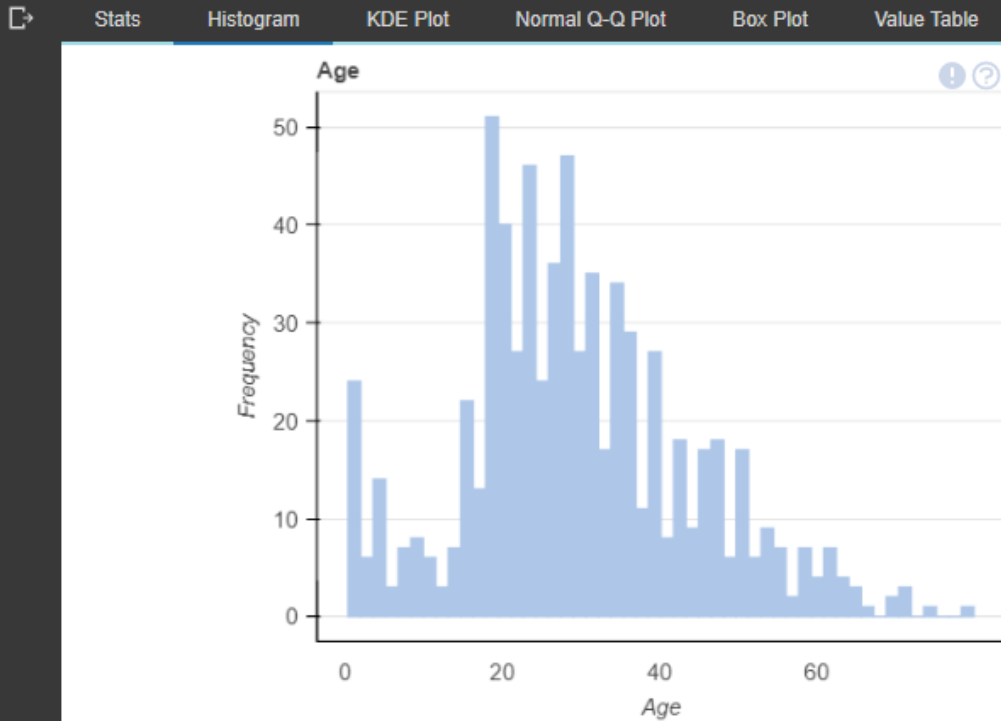
```
plot_correlation(df, "Survived")
```

```
#understand how other columns correlate to "Survived" column
```



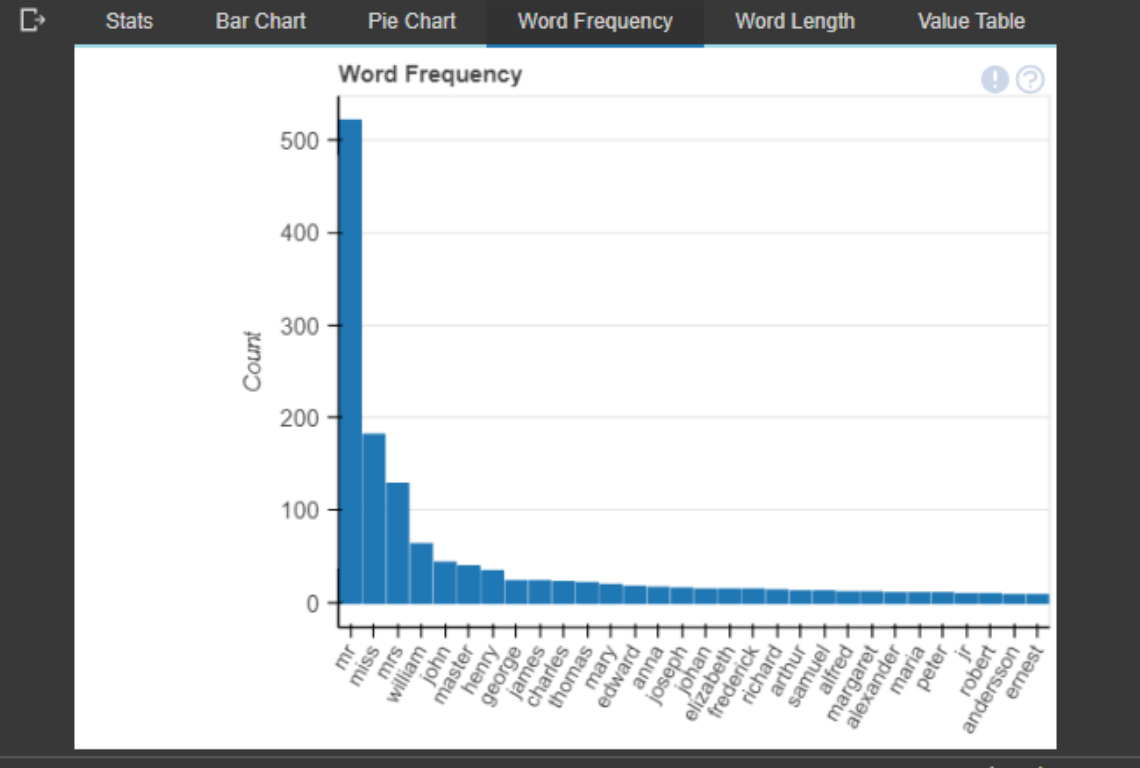
7) DataPrep - Describing a Single (Numerical) Column

```
plot(df, "Age")
```



8) DataPrep - Describing a Single (Categorical) Column

```
plot(df, "Name")
```



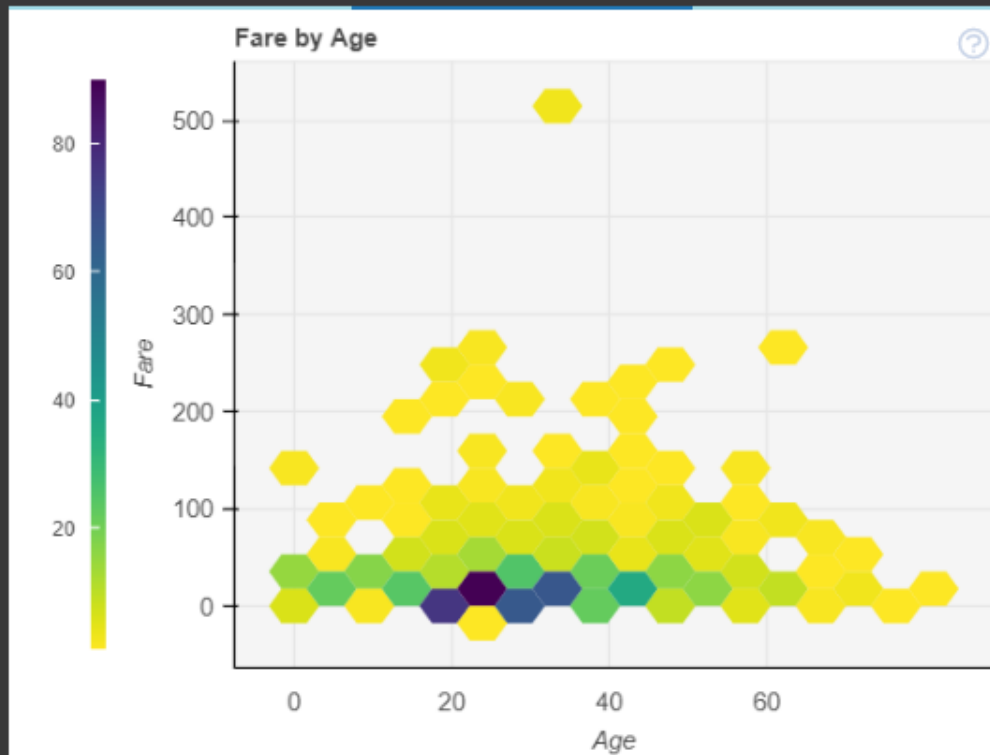
9) DataPrep - Describing Two Numerical Columns

```
plot(df, "Age", "Fare")
```

Scatter Plot

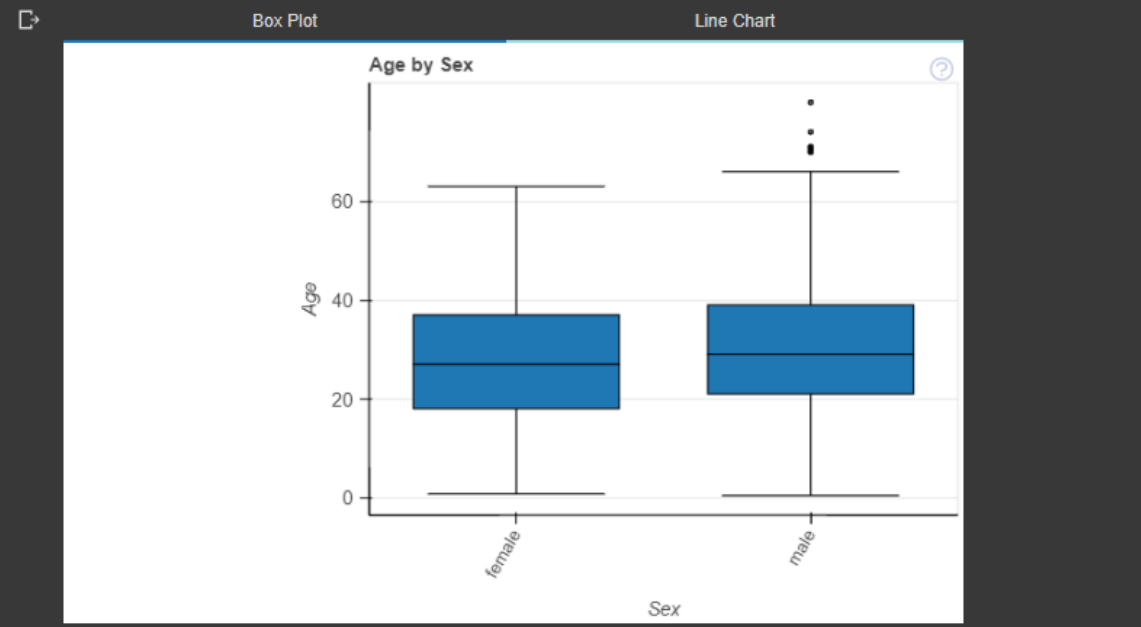
Hexbin Plot

Box Plot



10) DataPrep - Describing Two Columns (Numerical vs Categorical)

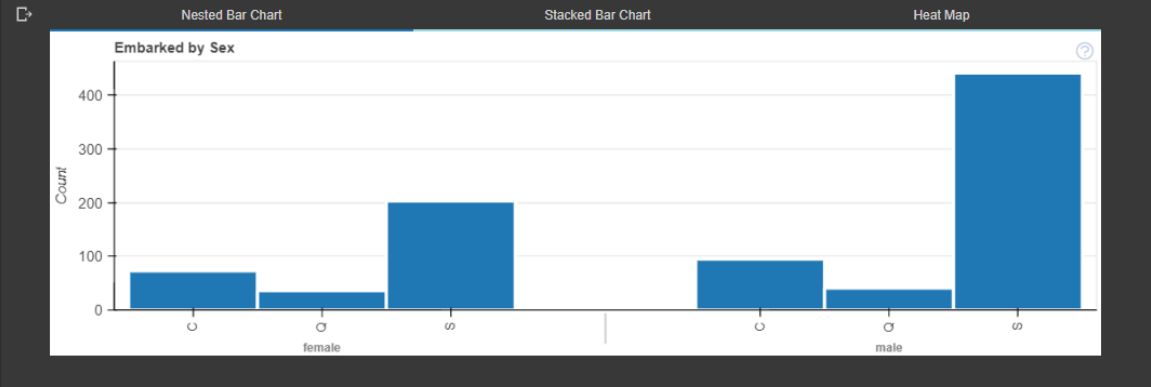
```
plot(df, "Age", "Sex")
```



XI. DATAPREP – DESCRIBING TWO CATEGORIAL COLUMNS

11) DataPrep - Describing Two Categorical Columns

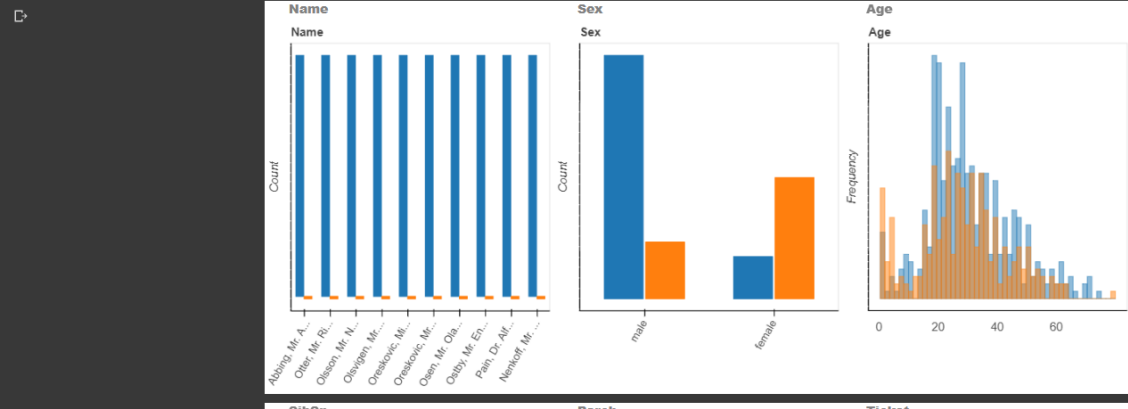
```
plot(df, "Sex", "Embarked")
```



XII. DATAPREP – COMPARE TWO DATAFRAMES

12) DataPrep - Compare Two Dataframes

```
df1 = df[df["Survived"] == 0]
df2 = df[df["Survived"] == 1]
plot_diff(df1, df2)
```

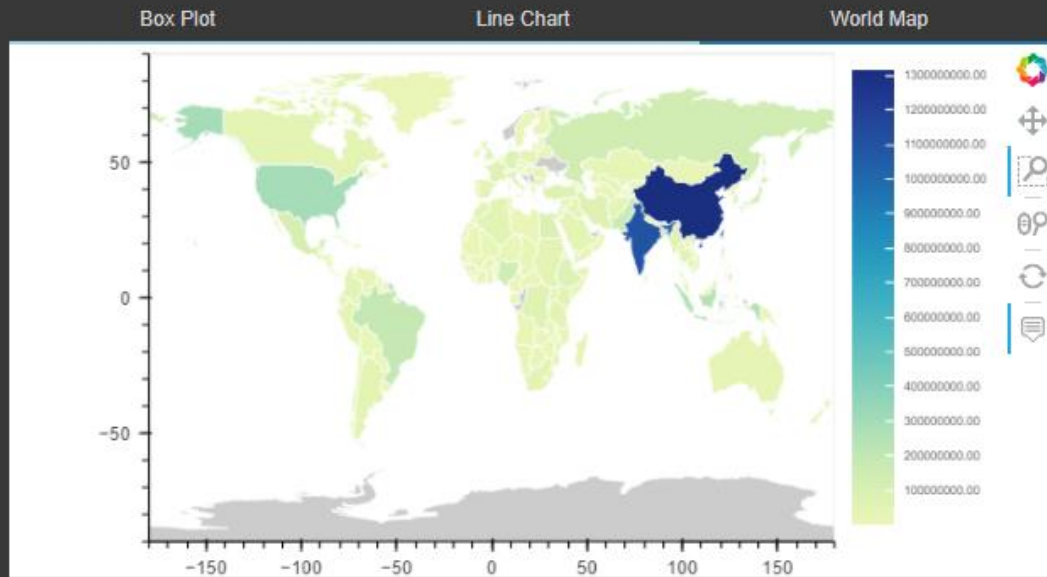


13) DataPrep - Create HTML Report

```
[ ] create_report(df).show_browser()
```

14) DataPrep - Analyze Geographical Data

```
[ ] country = load_dataset("countries")  
plot(country, "Country", "Population")
```



THE END

ABOUT DR. ALVIN ANG



Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.