

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=92HVlvi_6AI0

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text

0s

```
import pandas as pd

# Load dataset
url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
df = pd.read_csv(url)

# Display first few rows
print(df.head())

# Basic info
print(df.info())

# Check for null values
print(df.isnull().sum())
```

	PassengerId	Survived	Pclass	
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Sex	Age	SibSp	
0	Braund, Mr. Owen Harris	male	22.0	1	
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	
2	Heikkinen, Miss. Laina	female	26.0	0	

Release notes

Cell

```
# Fill missing 'Age' with median
df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing 'Embarked' with mode
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# Drop 'Cabin' due to too many missing values
df.drop(columns=['Cabin'], inplace=True)

# Verify
print(df.isnull().sum())
```

Variables

Terminal

Activate Windows

Go to Settings to activate Windows.

21:29 Python 3

21:30 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=92HVlvl_6Al0&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

↑ ↓ ↗ 🔗 💬 ⚙️ 📄 🗑️ ⋮

↩️ ↪️ ↺ ↻

PassengerId Survived Pclass \

0 1 0 3

1 2 1 1

2 3 1 3

3 4 1 1

4 5 0 3

Name Sex Age SibSp \

Braund, Mr. Owen Harris male 22.0 1

Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1

Heikkinen, Miss. Laina female 26.0 0

Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1

Allen, Mr. William Henry male 35.0 0

Parch Ticket Fare Cabin Embarked

0 0 A/5 21171 7.2500 NaN S

1 0 PC 17599 71.2833 C85 C

2 0 STON/O2. 3101282 7.9250 NaN S

3 0 113803 53.1000 C123 S

4 0 373450 8.0500 NaN S

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

Column Non-Null Count Dtype

0 PassengerId 891 non-null int64

1 Survived 891 non-null int64

2 Pclass 891 non-null int64

3 Name 891 non-null object

4 Sex 891 non-null object

5 Age 714 non-null float64

6 SibSp 891 non-null int64

7 Parch 891 non-null int64

8 Ticket 891 non-null object

9 Fare 891 non-null float64

10 Cabin 204 non-null object

11 Embarked 889 non-null object

dtypes: float64(2), int64(5), object(5)

memory usage: 83.7+ KB

None

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Release notes

Cell

Cell X

▶ ↩️ 📄

Data columns (total 12 columns):

Column Non-Null Count Dtype

0 PassengerId 891 non-null int64

1 Survived 891 non-null int64

2 Pclass 891 non-null int64

3 Name 891 non-null object

4 Sex 891 non-null object

5 Age 714 non-null float64

6 SibSp 891 non-null int64

7 Parch 891 non-null int64

8 Ticket 891 non-null object

9 Fare 891 non-null float64

10 Cabin 204 non-null object

11 Embarked 889 non-null object

dtypes: float64(2), int64(5), object(5)

memory usage: 83.7+ KB

None

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Variables Terminal

21:29 Python 3

Type here to search

32°C Haze 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsI4Pn4c-0T#scrollTo=92HVlvi_6AI0&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

<<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

Column Non-Null Count Dtype

0 PassengerId 891 non-null int64

1 Survived 891 non-null int64

2 Pclass 891 non-null int64

3 Name 891 non-null object

4 Sex 891 non-null object

5 Age 714 non-null float64

6 SibSp 891 non-null int64

7 Parch 891 non-null int64

8 Ticket 891 non-null object

9 Fare 891 non-null float64

10 Cabin 204 non-null object

11 Embarked 889 non-null object

dtypes: float64(2), int64(5), object(5)

memory usage: 83.7+ KB

None

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

dtype: int64

Release notes Cell Cell X

import pandas as pd

Load dataset

url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic"

df = pd.read_csv(url)

Display first few rows

print(df.head())

Basic info

print(df.info())

Check for null values

print(df.isnull().sum())

PassengerId Survived Pclass \

0 1 0 3

1 2 1 1

2 3 1 3

3 4 1 1

4 5 0 3

Name Sex Age SibSp \

0 Braund, Mr. Owen Harris male 22.0 1

1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1

2 Heikkinen, Miss. Laina female 26.0 0

Variables Terminal

21:29 Python 3

Type here to search

32°C Haze 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=X80Xhjc06Nch&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

Share Gemini

Release notes Cell X Cell

```
# Fill missing 'Age' with median
df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing 'Embarked' with mode
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# Drop 'Cabin' due to too many missing values
df.drop(columns=['Cabin'], inplace=True)

# Verify
print(df.isnull().sum())
```

PassengerId 0
Survived 0
Pclass 0
Name 0
Sex 0
Age 0
SibSp 0
Parch 0
Ticket 0
Fare 0
Embarked 0
dtype: int64
<ipython-input-9-ee2512204c12>:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

Variables Terminal

21:35 Python 3

Type here to search

21:36 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=X80Xhjc06Nch&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

RAM Disk

Share Gemini

Release notes Cell X Cell

PassengerId 0
Survived 0
Pclass 0
Name 0
Sex 0
Age 0
SibSp 0
Parch 0
Ticket 0
Fare 0
Embarked 0
dtype: int64

<ipython-input-9-ee2512204c12>:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method. The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation in

df['Age'].fillna(df['Age'].median(), inplace=True)
<ipython-input-9-ee2512204c12>:5: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method. The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation in

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

Activate Windows
Go to Settings to activate Windows.

Variables Terminal

21:35 Python 3

Type here to search

32°C Haze

21:36
26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=aTkFAy2lQjP_&uniqfier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

Convert 'Sex' and 'Embarked' using Label Encoding or One-Hot Encoding
df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=True)

Display updated dataframe
print(df.head())

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Age	SibSp	Parch	\
0	Braund, Mr. Owen Harris	22.0	1	0	
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1	0	
2	Heikkinen, Miss. Laina	26.0	0	0	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	1	0	
4	Allen, Mr. William Henry	35.0	0	0	

	Ticket	Fare	Sex_male	Embarked_Q	Embarked_S
0	A/5 21171	7.2500	True	False	True
1	PC 17599	71.2833	False	False	False
2	STON/O2. 3101282	7.9250	False	False	True
3	113803	53.1000	False	False	True
4	373450	8.0500	True	False	True

Release notes

Cell

Cell

Cell X

Convert 'Sex' and 'Embarked' using Label Encoding or One-Hot
df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=True)

Display updated dataframe
print(df.head())

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	

	Name	Age	SibSp
0	Braund, Mr. Owen Harris	22.0	1
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1
2	Heikkinen, Miss. Laina	26.0	0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	1
4	Allen, Mr. William Henry	35.0	0

	Ticket	Fare	Sex_male	Embarked_Q	Embarked_S
0	A/5 21171	7.2500	True	False	True
1	PC 17599	71.2833	False	False	False
2	STON/O2. 3101282	7.9250	False	False	True
3	113803	53.1000	False	False	True
4	373450	8.0500	True	False	True

Variables

Terminal

21:37 Python 3

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=TCaCY8FJUAgO&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

1 PC 17599 71.2833 False False False

2 STON/O2. 3101282 7.9250 False False True

3 113803 53.1000 False False True

4 373450 8.0500 True False True

from sklearn.preprocessing import StandardScaler

Select numerical columns

numerical_cols = ['Age', 'Fare', 'SibSp', 'Parch']

Initialize scaler

scaler = StandardScaler()

df[numerical_cols] = scaler.fit_transform(df[numerical_cols])

Display scaled data

print(df[numerical_cols].head())

Age Fare SibSp Parch

0 -0.565736 -0.502445 0.432793 -0.473674

1 0.663861 0.786845 0.432793 -0.473674

2 -0.258337 -0.488854 -0.474545 -0.473674

3 0.433312 0.420730 0.432793 -0.473674

4 0.433312 -0.486337 -0.474545 -0.473674

[] import seaborn as sns

import matplotlib.pyplot as plt

Plot boxplots

for col in numerical_cols:

Release notes Cell Cell Cell X

Convert 'Sex' and 'Embarked' using Label Encoding or One-Hot I

df = pd.get_dummies(df, columns=['Sex', 'Embarked'], drop_first=

Display updated dataframe

print(df.head())

PassengerId Survived Pclass \

0 1 0 3

1 2 1 1

2 3 1 3

3 4 1 1

4 5 0 3

Name Age SibSp

0 Braund, Mr. Owen Harris 22.0 1

1 Cumings, Mrs. John Bradley (Florence Briggs Th... 38.0 1

2 Heikkinen, Miss. Laina 26.0 0

3 Futrelle, Mrs. Jacques Heath (Lily May Peel) 35.0 1

4 Allen, Mr. William Henry 35.0 0

Ticket Fare Sex_male Embarked_Q Embarked_S

0 A/5 21171 7.2500 True False True

1 PC 17599 71.2833 False False False

2 STON/O2. 3101282 7.9250 False False True

3 113803 53.1000 False False True

4 373450 8.0500 True False True

Variables Terminal

21:39 Python 3

Type here to search

32°C Haze 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFhELD9nYe323qEsl4Pn4c-0T#scrollTo=XuUtZVCbUJZy&uniqifier=1

Untitled0.ipynb

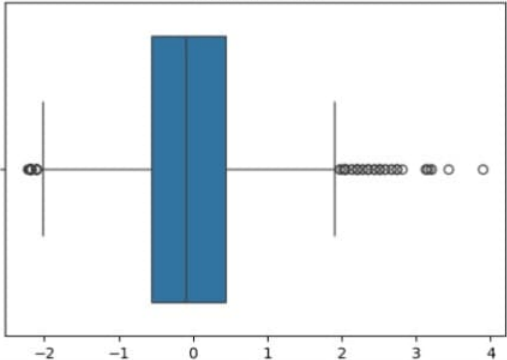
File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

```
import seaborn as sns
import matplotlib.pyplot as plt

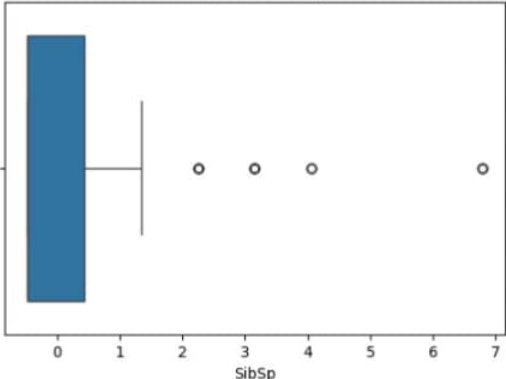
# Plot boxplots
for col in numerical_cols:
    plt.figure(figsize=(6, 4))
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
```

Boxplot of Age



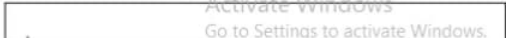
A boxplot showing the distribution of Age. The x-axis ranges from -2 to 4. The box is centered around 0, with whiskers extending from approximately -1.8 to 1.8. There are several outliers at higher values, around 2.5, 3, and 4.

Boxplot of SibSp



A boxplot showing the distribution of SibSp. The x-axis ranges from 0 to 7. The box is centered around 0, with whiskers extending from approximately -0.5 to 1.5. There are several outliers at higher values, around 2.5, 3, 4, and 6.5.

Boxplot of Parch



A boxplot showing the distribution of Parch. The x-axis ranges from 0 to 7. The box is centered around 0, with whiskers extending from approximately -0.5 to 1.5. There are several outliers at higher values, around 2.5, 3, 4, and 6.5.

Variables

Terminal

21:39 Python 3

Type here to search

32°C Haze

21:40 26-05-2025

Colab

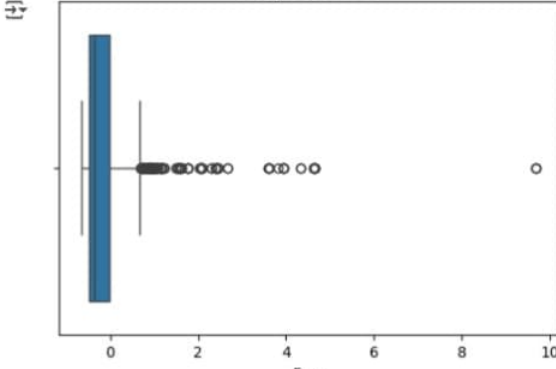
colab.research.google.com/drive/1ngS0dOtKF...#scrollTo=XuUtZVCbUJZy&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help


Commands Code Text

Boxplot of Fare



A boxplot titled 'Boxplot of Fare' showing the distribution of fare values. The x-axis is labeled 'Fare' and ranges from 0 to 10. The plot shows a median near 0, a box from approximately -0.5 to 0.5, whiskers from -1 to 1, and several outliers between 1 and 10.

Boxplot of SibSp

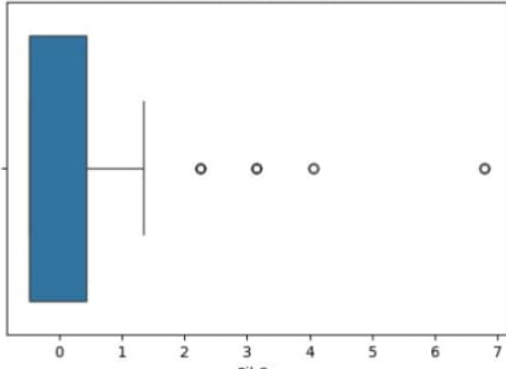


A boxplot titled 'Boxplot of SibSp' showing the distribution of the number of siblings/spouses aboard. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The plot shows a median at 0, a box from approximately -0.5 to 0.5, and whiskers from -1 to 1.5. There are no outliers.

Release notes

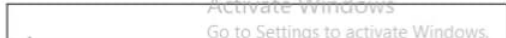
Cell Cell Cell Cell X

Boxplot of SibSp



A boxplot titled 'Boxplot of SibSp' showing the distribution of the number of siblings/spouses aboard. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The plot shows a median at 0, a box from approximately -0.5 to 0.5, and whiskers from -1 to 1.5. There are no outliers.

Boxplot of Parch



A boxplot titled 'Boxplot of Parch' showing the distribution of the number of parents/children aboard. The x-axis is labeled 'Parch' and ranges from 0 to 7. The plot shows a median at 0, a box from approximately -0.5 to 0.5, and whiskers from -1 to 1.5. There are no outliers.

Variables Terminal

21:39 Python 3

Type here to search

32°C Haze

26-05-2025

Untitled0.ipynb - Colab

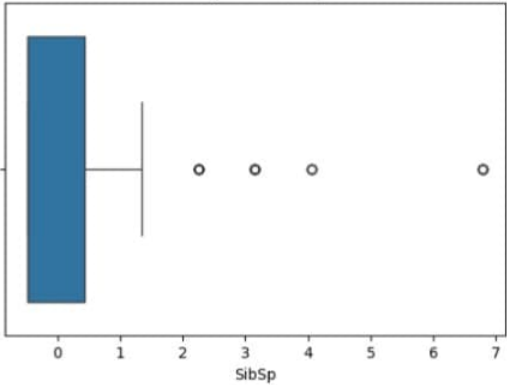
colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsl4Pn4c-0T#scrollTo=XuUtZVCbUJZy&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text


Boxplot of SibSp



A boxplot showing the distribution of SibSp (Siblings/Spouse) for the Fare variable. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The y-axis is labeled 'Fare'. The boxplot shows a median around 0.5, with a box from approximately 0.1 to 0.9. Whiskers extend from 0 to 1.5. There are four outliers at approximately 2.2, 3.0, 4.0, and 6.8.

SibSp	Fare
0	~5.4
1	~3.7
2	~12.0
3	~12.0
4	~12.0
6	~53.0

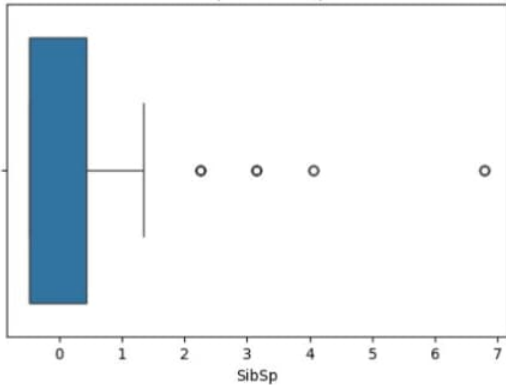
Boxplot of Parch



A boxplot showing the distribution of Parch (Parents/Children) for the Fare variable. The x-axis is labeled 'Parch' and ranges from 0 to 1. The y-axis is labeled 'Fare'. The boxplot shows a median around 0.5, with a box from approximately 0.1 to 0.9. Whiskers extend from 0 to 1. There are no outliers.

Parch	Fare
0	~5.4
1	~3.7

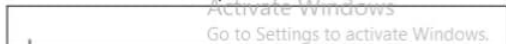
Boxplot of SibSp



A boxplot showing the distribution of SibSp (Siblings/Spouse) for the Fare variable. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The y-axis is labeled 'Fare'. The boxplot shows a median around 0.5, with a box from approximately 0.1 to 0.9. Whiskers extend from 0 to 1.5. There are four outliers at approximately 2.2, 3.0, 4.0, and 6.8.

SibSp	Fare
0	~5.4
1	~3.7
2	~12.0
3	~12.0
4	~12.0
6	~53.0

Boxplot of Parch



A boxplot showing the distribution of Parch (Parents/Children) for the Fare variable. The x-axis is labeled 'Parch' and ranges from 0 to 1. The y-axis is labeled 'Fare'. The boxplot shows a median around 0.5, with a box from approximately 0.1 to 0.9. Whiskers extend from 0 to 1. There are no outliers.

Parch	Fare
0	~5.4
1	~3.7

Variables

Terminal

21:39

Python 3

Type here to search

32°C Haze

26-05-2025

Untitled0.ipynb - Colab

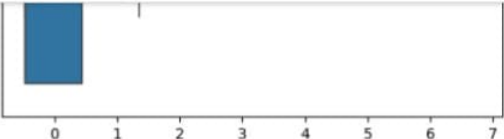
colab.research.google.com/drive/1ngS0dOtKFhELD9nYe323qEsl4Pn4c-0T#scrollTo=XuUtZVCbUJZy&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

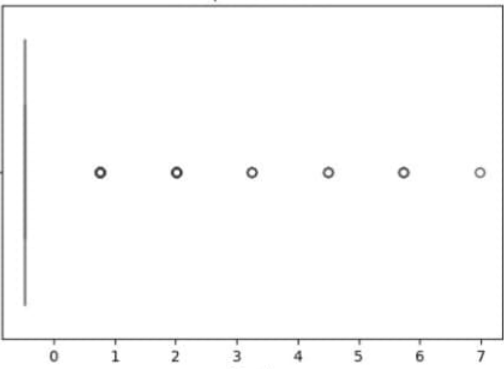
Commands + Code + Text

Boxplot of SibSp



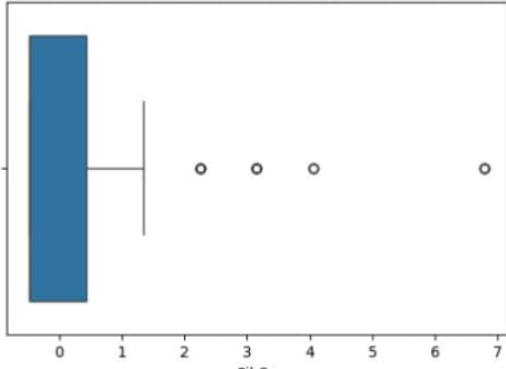
A boxplot showing the distribution of SibSp (Siblings/Spouse) values. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The plot shows a single box from 0 to 1, with a median line at 0.5. There are no outliers.

Boxplot of Parch



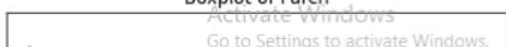
A boxplot showing the distribution of Parch (Parents/Children) values. The x-axis is labeled 'Parch' and ranges from 0 to 7. The plot shows a single box from 0 to 1, with a median line at 0.5. There are no outliers.

Boxplot of SibSp



A boxplot showing the distribution of SibSp (Siblings/Spouse) values. The x-axis is labeled 'SibSp' and ranges from 0 to 7. The plot shows a box from 0 to 1, with a median line at 0.5. There are four outliers at 2, 3, 4, and 7.

Boxplot of Parch



A boxplot showing the distribution of Parch (Parents/Children) values. The x-axis is labeled 'Parch' and ranges from 0 to 7. The plot shows a box from 0 to 1, with a median line at 0.5. There are four outliers at 2, 3, 4, and 7.

Variables

Terminal

21:39 Python 3

Type here to search

32°C Haze 26-05-2025

Untitled0.ipynb - Colab

colab.research.google.com/drive/1ngS0dOtKFnELD9nYe323qEsI4Pn4c-0T#scrollTo=Np92pNBFUgyF&uniqifier=1

Untitled0.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

0 1 2 3 4 5 6 7

Parch

Function to remove outliers

def remove_outliers(df, column):

Q1 = df[column].quantile(0.25)

Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1

lower = Q1 - 1.5 * IQR

upper = Q3 + 1.5 * IQR

return df[(df[column] >= lower) & (df[column] <= upper)]

Apply outlier removal on all numerical columns

for col in numerical_cols:

df = remove_outliers(df, col)

Final shape after removing outliers

print(df.shape)

(577, 12)

Release notes

Cell Cell Cell Cell X

Fare

Boxplot of SibSp

0 1 2 3 4 5 6 7

SibSp

Boxplot of Parch

Activate Windows

Go to Settings to activate Windows.

Variables Terminal

21:44 Python 3

Type here to search

32°C Haze

21:44 26-05-2025