

Web Traffic Analysis



Web Traffic Analysis Phase-3



Presented by:

N.Soleswaran
V.Vinothan
R.Sri jaya shankaran
S.Thamizhselvan
Soalwin thomas

Web Traffic Analysis

Monitoring web traffic requires information such as the total number of visitors, average page views per visitor, most popular pages, average visits by visitors.

Remove duplicate observations

Duplicate data most often occurs during the data collection process. This typically happens when you combine data from multiple places, or receive data from clients or multiple departments.

Filter unwanted outliers

Outliers are unusual values in your dataset. They're significantly different from other data point and can distort your analysis and violate assumptions. Removing them is a subjective practice and depends on what you're trying to analyze.

Fix structural errors

Structural errors are things like strange naming conventions, typos, or incorrect capitalization. Anything that is inconsistent will create mislabeled categories.

The pandas have been installed into the system, we need to import the library

Seaborn is a library for making statistical graphics in Python

The screenshot shows a Jupyter Notebook interface with the following code cells and outputs:

```
In [2]: import pandas as pd
import seaborn as sns
```

```
In [3]: sns.set(color_codes=True)
```

```
In [4]: project = pd.read_csv('daily-website-visitors.csv')
```

```
In [5]: project.head()
```

Out[5]:

	Row	Day	Day.Of.Week	Date	Page.Loads	Unique.Visits	First.Time.Visits	Returning.Visits
0	1	Sunday	1	9/14/2014	2,146	1,582	1,430	152
1	2	Monday	2	9/15/2014	3,621	2,528	2,297	231
2	3	Tuesday	3	9/16/2014	3,698	2,630	2,352	278
3	4	Wednesday	4	9/17/2014	3,667	2,614	2,327	287
4	5	Thursday	5	9/18/2014	3,316	2,366	2,130	236

```
In [6]: project.tail()
```

Out[6]:

	Row	Day	Day.Of.Week	Date	Page.Loads	Unique.Visits	First.Time.Visits	Returning.Visits
2162	2163	Saturday	7	8/15/2020	2,221	1,696	1,373	323
2163	2164	Sunday	1	8/16/2020	2,724	2,037	1,686	351
2164	2165	Monday	2	8/17/2020	3,456	2,638	2,181	457
2165	2166	Tuesday	3	8/18/2020	3,581	2,683	2,184	499
2166	2167	Wednesday	4	8/19/2020	2,064	1,564	1,297	267

```
In [8]: project.info()
```

The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column

Reminder to join - IBM NaanMu x IBM NaanMudhalvan - Data Ana x Untitled Folder/ x Untitled - Jupyter Notebook x WhatsApp x

localhost:8888/notebooks/Untitled%20Folder/Untitled.ipynb

Jupyter Untitled Last Checkpoint: 13 hours ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)


In [8]: `project.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2167 entries, 0 to 2166
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ---
 0   Row                    2167 non-null   int64
 1   Day                    2167 non-null   object
 2   Day.Of.Week            2167 non-null   int64
 3   Date                   2167 non-null   object
 4   Page.Loads             2167 non-null   object
 5   Unique.Visits          2167 non-null   object
 6   First.Time.Visits      2167 non-null   object
 7   Returning.Visits       2167 non-null   object
dtypes: int64(2), object(6)
memory usage: 135.6+ KB
```

In [17]: `sns.barplot(project['Row'], project['Date'])`

C:\Users\solesh\anaconda3\lib\site-packages\seaborn_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

Out[17]: <AxesSubplot:xlabel='Row', ylabel='Date'>



The `dropna()` method returns a new `inplace` parameter is set to `True`, in that case the `dropna` method does the removing in the original DataFrame instead.

Use the `subset` parameter if only some specified columns should be considered when looking for duplicates.

After completing the process Extract the cleaned data set then move to the next visualization



jupyter Untitled Last Checkpoint: 13 hours ago (unsaved changes)



Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

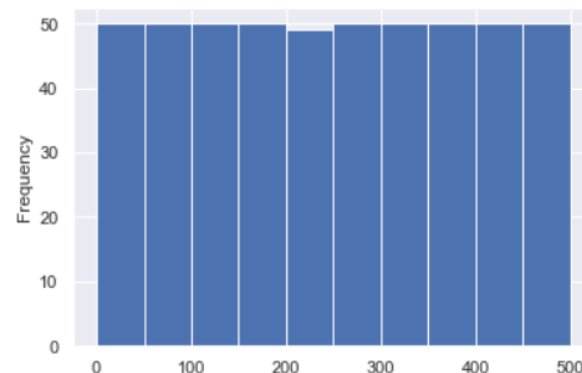


Python 3 (ipykernel)

Run Code

```
In [38]: project["Row"].plot(kind = 'hist')
```

```
Out[38]: <AxesSubplot:ylabel='Frequency'>
```



```
In [5]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [6]: df = pd.read_csv('webtraffic-analysis.csv')
```

```
In [7]: df = df.dropna()
```

```
In [8]: df = df.drop_duplicates()
```

```
In [14]: df.to_csv('webtraffic-analysis.csv',index=False)
```

```
In [ ]:
```



Type here to search



32°C Haze



12:47 PM
10/18/2023





Thank You



WEB TRAFFIC

iStock
Credit: relif