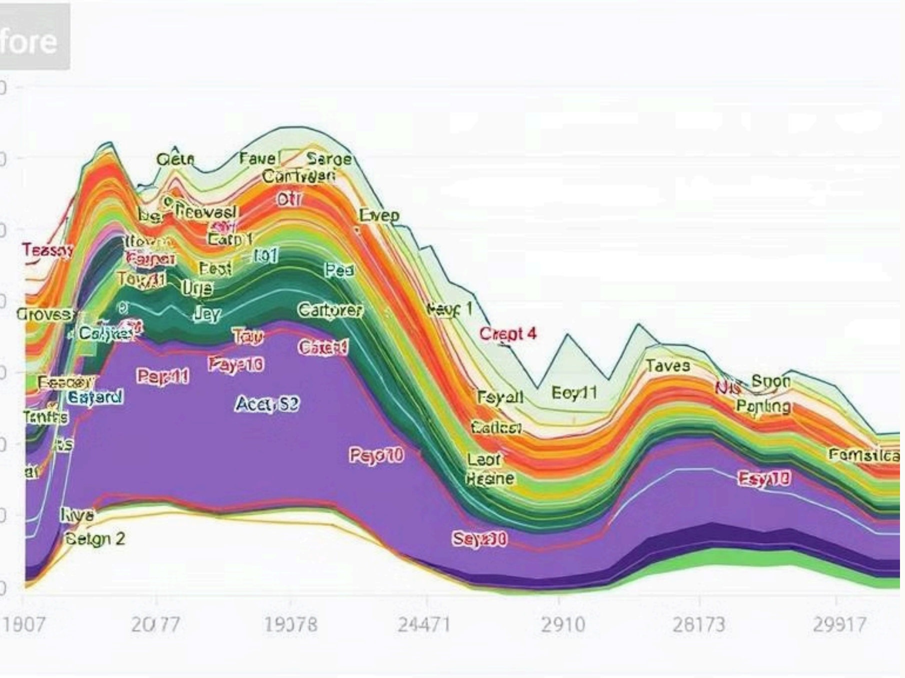


Data Harmonization and Insights Extraction

Discover how harmonizing data unlocks valuable insights for improved decision-making and competitive advantage.

by Priyanka Mathur

Data Harmonization using tools and techniques

Techniques: o

Merging datasets. o Imputation of

Outlier detection and handling. o

Data validation and exploration.

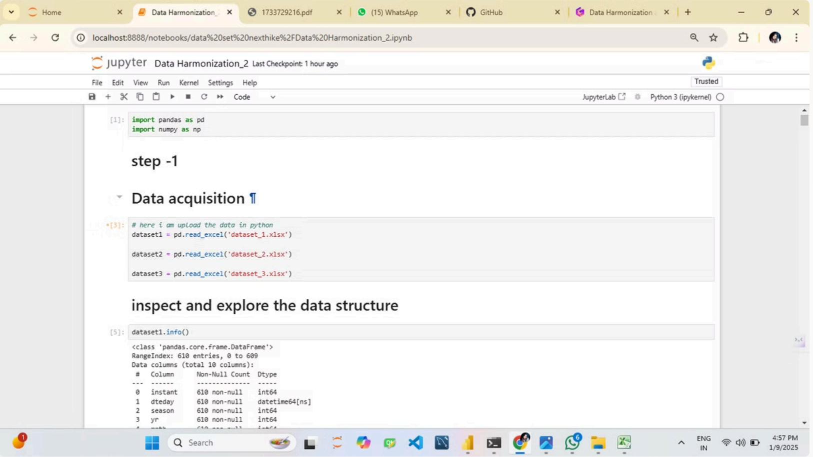
Tools: o

Pandas for data manipulation.

NumPy for numerical computations. •.

1: Data Acquisition

- Load datasets into Python.
- Inspect and explore data structures using .info() and .head().
- Document initial observations about data quality and completeness.


Home
Data Harmonization_
1733729216.pdf
(15) WhatsApp
GitHub
Data Harmonization
localhost:8888/notebooks/data%20set%20nexthike%2FData%20Harmonization_2.ipynb
jupyter Data Harmonization_2 Last Checkpoint: 1 hour ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
Search
ENG IN
4:37 PM 1/9/2025

```
[1]: import pandas as pd
     import numpy as np
```

step -1

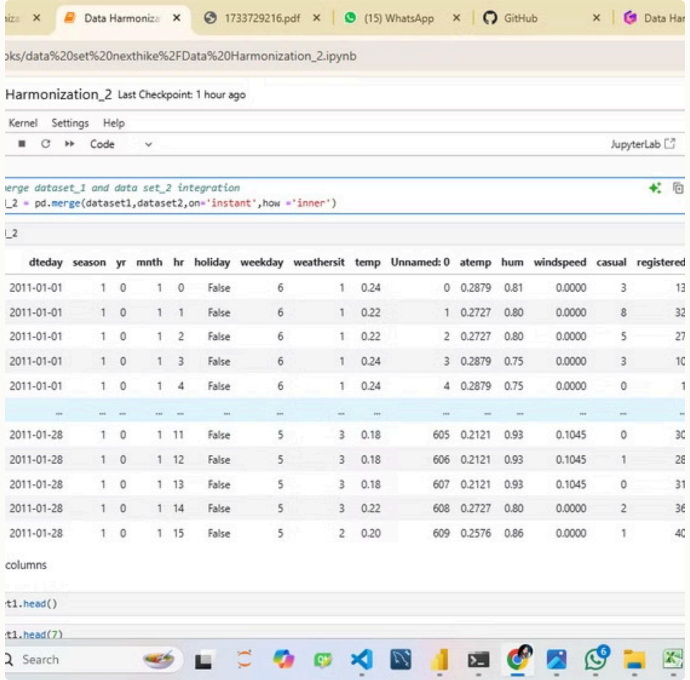
Data acquisition ¶

```
[3]: # here i am upload the data in python
     dataset1 = pd.read_excel('dataset_1.xlsx')

     dataset2 = pd.read_excel('dataset_2.xlsx')

     dataset3 = pd.read_excel('dataset_3.xlsx')
```

inspect and explore the data structure

```
[5]: dataset1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 610 entries, 0 to 609
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   instant     610 non-null    int64
 1   dteday      610 non-null    datetime64[ns]
 2   season      610 non-null    int64
 3   yr          610 non-null    int64
```

- # Dataset_3 Integration


Data Harmoniza
1733729216.pdf
(15) WhatsApp
GitHub
Data Har
oks/data%20set%20nexthike%2FData%20Harmonization_2.ipynb
Harmonization_2 Last Checkpoint: 1 hour ago
Kernel
Settings
Help
Code
JupyterLab
Search

```
erge dataset_1 and data set_2 integration
_2 = pd.merge(dataset1,dataset2,on='instant',how ='inner')
```

```
_2
```

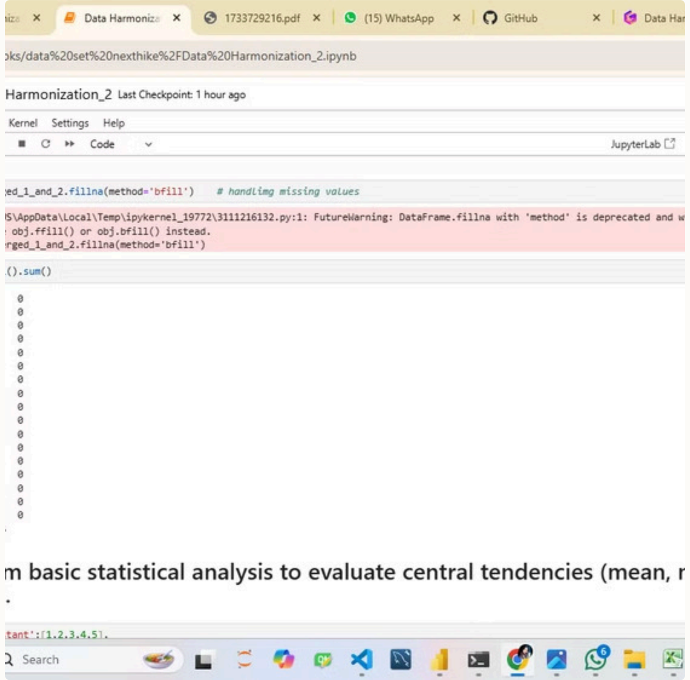
dteday	season	yr	mnth	hr	holiday	weekday	weathersit	temp	Unnamed: 0	atemp	hum	windspeed	casual	registered
2011-01-01	1	0	1	0	False	6	1	0.24	0	0.2879	0.81	0.0000	3	13
2011-01-01	1	0	1	1	False	6	1	0.22	1	0.2727	0.80	0.0000	8	32
2011-01-01	1	0	1	2	False	6	1	0.22	2	0.2727	0.80	0.0000	5	27
2011-01-01	1	0	1	3	False	6	1	0.24	3	0.2879	0.75	0.0000	3	10
2011-01-01	1	0	1	4	False	6	1	0.24	4	0.2879	0.75	0.0000	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2011-01-28	1	0	1	11	False	5	3	0.18	605	0.2121	0.93	0.1045	0	30
2011-01-28	1	0	1	12	False	5	3	0.18	606	0.2121	0.93	0.1045	1	28
2011-01-28	1	0	1	13	False	5	3	0.18	607	0.2121	0.93	0.1045	0	31
2011-01-28	1	0	1	14	False	5	3	0.22	608	0.2727	0.80	0.0000	2	36
2011-01-28	1	0	1	15	False	5	2	0.20	609	0.2576	0.86	0.0000	1	40

columns

```
t1.head()
```

```
t1.head(7)
```

data set integration


Data Harmoniza
1733729216.pdf
(15) WhatsApp
GitHub
Data Har
oks/data%20set%20nexthike%2FData%20Harmonization_2.ipynb
Harmonization_2 Last Checkpoint: 1 hour ago
Kernel
Settings
Help
Code
JupyterLab
Search

```
ed_1_and_2.fillna(method='bfill')    # handlimg missing values
```

```
S\AppData\Local\Temp\ipykernel_19772\3111216132.py:1: FutureWarning: DataFrame.fillna with 'method' is deprecated and w
 obj.ffill() or obj.bfill() instead.
rged_1_and_2.fillna(method='bfill')
```

```
().sum()
```

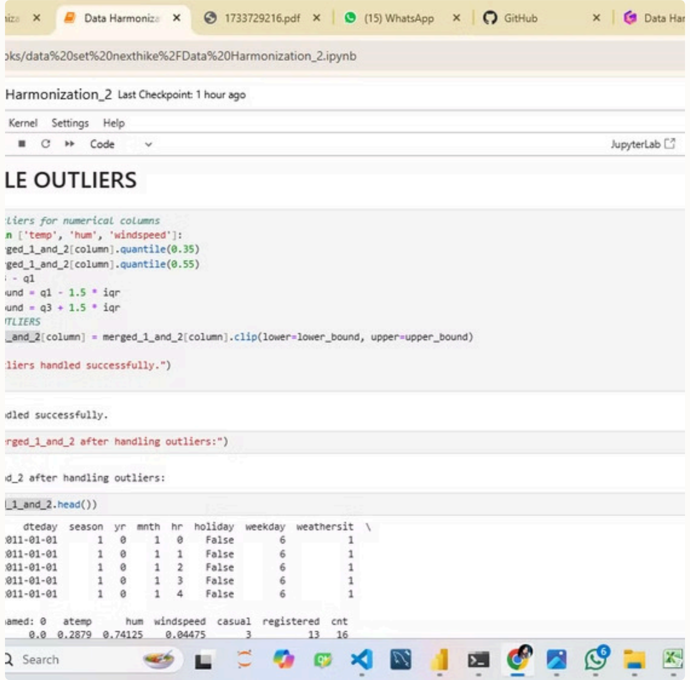
```
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
 0
```

m basic statistical analysis to evaluate central tendencies (mean, r

.

```
tant':[1,2,3,4,5],
```

handling missing value


Data Harmoniza
1733729216.pdf
(15) WhatsApp
GitHub
Data Har
oks/data%20set%20nexthike%2FData%20Harmonization_2.ipynb
Harmonization_2 Last Checkpoint: 1 hour ago
Kernel
Settings
Help
Code
JupyterLab
Search

LE OUTLIERS

```
liers for numerical columns
n ['temp', 'hum', 'windspeed']:
ged_1_and_2[column].quantile(0.35)
ged_1_and_2[column].quantile(0.55)
 - q1
und = q1 - 1.5 * iqr
und = q3 + 1.5 * iqr
TLIERS
_and_2[column] = merged_1_and_2[column].clip(lower=lower_bound, upper=upper_bound)

liers handled successfully.")
```

```
dled successfully.
```

```
rged_1_and_2 after handling outliers:")
```

```
d_2 after handling outliers:
```

```
_1_and_2.head())
```

```
   dteday  season  yr  mnth  hr  holiday  weekday  weathersit  \
011-01-01       1   0     1   0    False        6           1
011-01-01       1   0     1   1    False        6           1
011-01-01       1   0     1   2    False        6           1
011-01-01       1   0     1   3    False        6           1
011-01-01       1   0     1   4    False        6           1

amed: 0   atemp      hum  windspeed  casual  registered  cnt
    0.0  0.2879  0.74125    0.04475       3          13   16
```

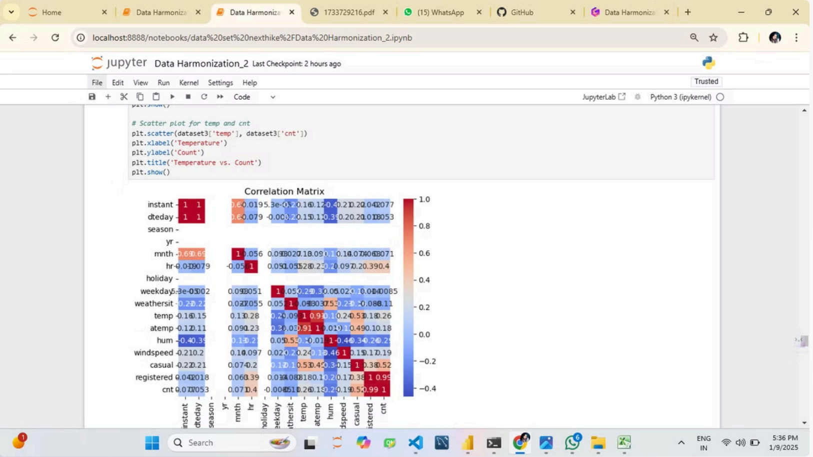
handle outliers

Made with Gamma

Compute correlations between attributes to identify relationships.

using matplolip

using seaborn



```
# Scatter plot for temp and cnt
plt.scatter(dataset3['temp'], dataset3['cnt'])
plt.xlabel('Temperature')
plt.ylabel('Count')
plt.title('Temperature vs. Count')
plt.show()
```

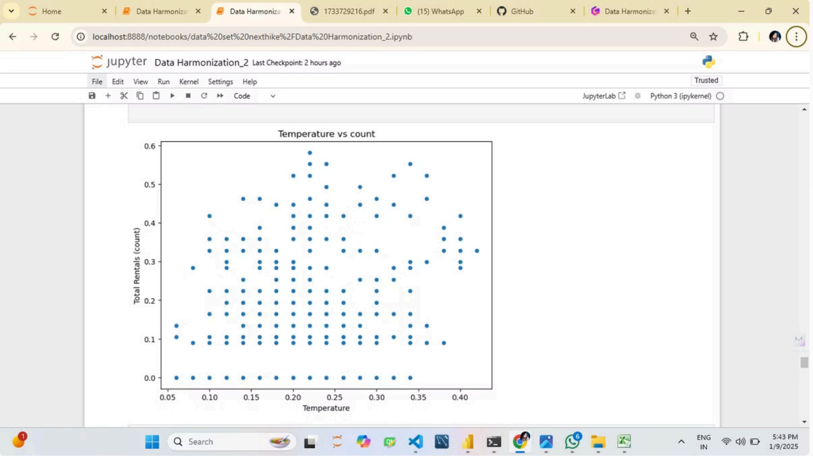
Made with Gamma

scatter plots

Visualizations

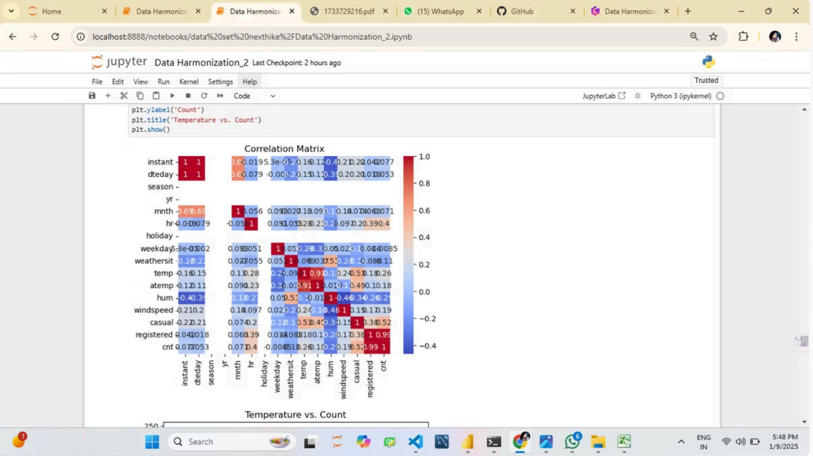
Scatterplot of Temperature vs Count

plt.figure(figsize=(8, 6)) sns.scatterplot(dataset3, x='temp', y='windspeed') plt.title("Temperature vs count") plt.xlabel("Temperature") plt.ylabel("Total Rentals (count)") plt.show()

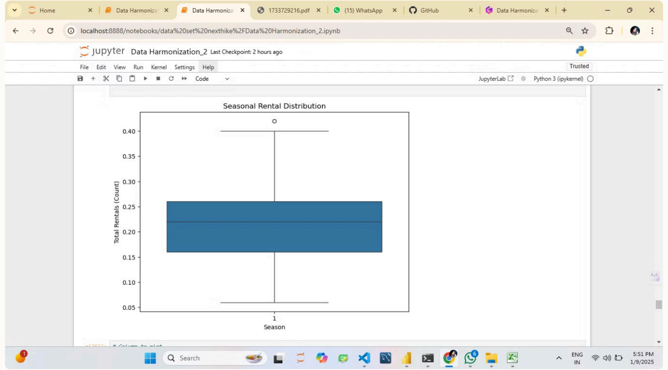


heapmap

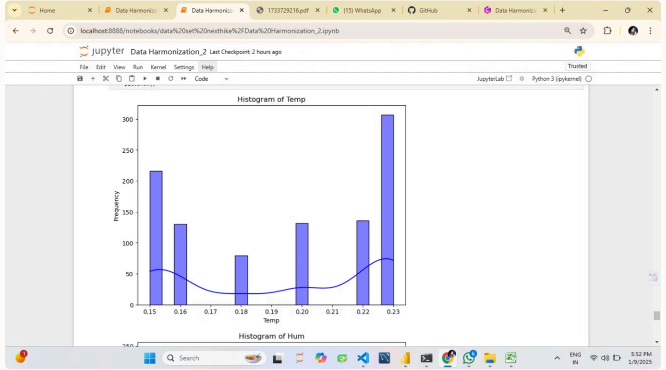
correlation_matrix = dataset3.corr() sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm') plt.title("Correlation Matrix") plt.show()



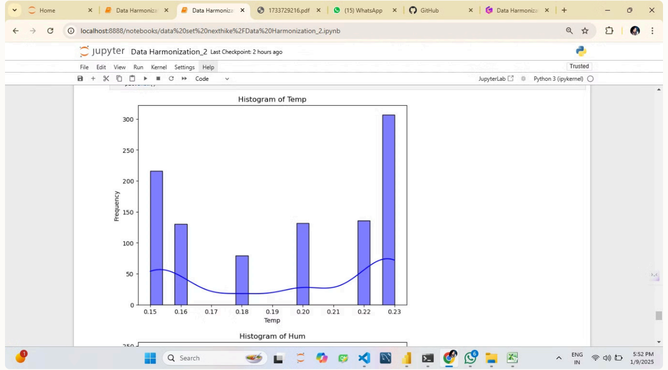
```
plt.ylabel('Count')
plt.title('Temperature vs. Count')
plt.show()
```



box plot



histogram



Histogram

Thank You