

# 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.

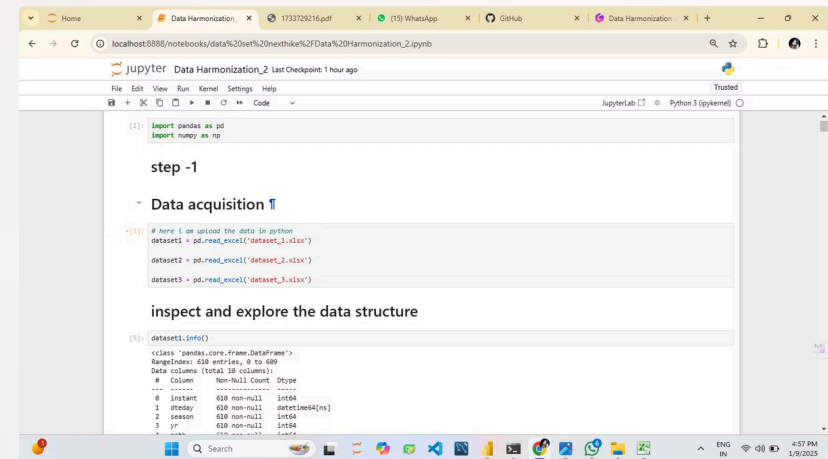
Tools: o

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.



The screenshot shows a JupyterLab notebook interface with a browser window. The notebook is titled "Data Harmonization\_2" and has a "Trusted" status. The code is organized into sections: "step -1", "Data acquisition", and "inspect and explore the data structure".

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

step -1

Data acquisition

# Here I use 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 dayofweek       610 non-null    int64[ms]
 2 season         610 non-null    int64
 3 yr             610 non-null    int64
 4 ...            ...
```

- Dataset\_3 Integration

```
merge dataset_1 and data set_2 integration
L2 = pd.merge(datasets1,dataset2,on='instant',how = 'inner')

L2

   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    33
2011-01-01    1    0    1    2    False        6         1    0.22         2    0.2727    0.80    0.0000         5    21
2011-01-01    1    0    1    3    False        6         1    0.24         3    0.2879    0.75    0.0000         3    16
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    36
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    33
2011-01-28    1    0    1   14    False        5         3    0.22        608    0.2727    0.80    0.0000         2         3
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

```
137: Data Harmonization_2.ipynb
138: 1733729216.pdf
139: (15) WhatsApp
140: GitHub
141: Data Harmonization_2.ipynb

Harmonization_2 Last Checkpoint: 1 hour ago

Kernel Settings Help

Code

JupyterLab

In [ ]:
1: red_1_and_2.fillna(method='bfill') # handling missing values
2:
3: #AppData\Local\Temp\ipykernel_19772\3111216132.py:1: FutureWarning: DataFrame.fillna with 'method' is deprecated and
4: will be removed in a future version. Use obj.ffill() or obj.bfill() instead.
5: red_1_and_2.fillna(method='bfill')
6:
7:
8:
9:
10:
11:
12:
13:
14:
15:
16:
17:
18:
19:
20:
21:
22:
23:
24:
25:
26:
27:
28:
29:
30:
31:
32:
33:
34:
35:
36:
37:
38:
39:
40:
41:
42:
43:
44:
45:
46:
47:
48:
49:
50:
51:
52:
53:
54:
55:
56:
57:
58:
59:
60:
61:
62:
63:
64:
65:
66:
67:
68:
69:
70:
71:
72:
73:
74:
75:
76:
77:
78:
79:
80:
81:
82:
83:
84:
85:
86:
87:
88:
89:
90:
91:
92:
93:
94:
95:
96:
97:
98:
99:
100:
101:
102:
103:
104:
105:
106:
107:
108:
109:
110:
111:
112:
113:
114:
115:
116:
117:
118:
119:
120:
121:
122:
123:
124:
125:
126:
127:
128:
129:
130:
131:
132:
133:
134:
135:
136:
137:
138:
139:
140:
141:
142:
143:
144:
145:
146:
147:
148:
149:
150:
151:
152:
153:
154:
155:
156:
157:
158:
159:
160:
161:
162:
163:
164:
165:
166:
167:
168:
169:
170:
171:
172:
173:
174:
175:
176:
177:
178:
179:
180:
181:
182:
183:
184:
185:
186:
187:
188:
189:
190:
191:
192:
193:
194:
195:
196:
197:
198:
199:
200:
201:
202:
203:
204:
205:
206:
207:
208:
209:
210:
211:
212:
213:
214:
215:
216:
217:
218:
219:
220:
221:
222:
223:
224:
225:
226:
227:
228:
229:
230:
231:
232:
233:
234:
235:
236:
237:
238:
239:
240:
241:
242:
243:
244:
245:
246:
247:
248:
249:
250:
251:
252:
253:
254:
255:
256:
257:
258:
259:
260:
261:
262:
263:
264:
265:
266:
267:
268:
269:
270:
271:
272:
273:
274:
275:
276:
277:
278:
279:
280:
281:
282:
283:
284:
285:
286:
287:
288:
289:
290:
291:
292:
293:
294:
295:
296:
297:
298:
299:
300:
301:
302:
303:
304:
305:
306:
307:
308:
309:
310:
311:
312:
313:
314:
315:
316:
317:
318:
319:
320:
321:
322:
323:
324:
325:
326:
327:
328:
329:
330:
331:
332:
333:
334:
335:
336:
337:
338:
339:
340:
341:
342:
343:
344:
345:
346:
347:
348:
349:
350:
351:
352:
353:
354:
355:
356:
357:
358:
359:
360:
361:
362:
363:
364:
365:
366:
367:
368:
369:
370:
371:
372:
373:
374:
375:
376:
377:
378:
379:
380:
381:
382:
383:
384:
385:
386:
387:
388:
389:
390:
391:
392:
393:
394:
395:
396:
397:
398:
399:
400:
401:
402:
403:
404:
405:
406:
407:
408:
409:
410:
411:
412:
413:
414:
415:
416:
417:
418:
419:
420:
421:
422:
423:
424:
425:
426:
427:
428:
429:
430:
431:
432:
433:
434:
435:
436:
437:
438:
439:
440:
441:
442:
443:
444:
445:
446:
447:
448:
449:
450:
451:
452:
453:
454:
455:
456:
457:
458:
459:
460:
461:
462:
463:
464:
465:
466:
467:
468:
469:
470:
471:
472:
473:
474:
475:
476:
477:
478:
479:
480:
481:
482:
483:
484:
485:
486:
487:
488:
489:
490:
491:
492:
493:
494:
495:
496:
497:
498:
499:
500:
501:
502:
503:
504:
505:
506:
507:
508:
509:
510:
511:
512:
513:
514:
515:
516:
517:
518:
519:
520:
521:
522:
523:
524:
525:
526:
527:
528:
529:
530:
531:
532:
533:
534:
535:
536:
537:
538:
539:
540:
541:
542:
543:
544:
545:
546:
547:
548:
549:
550:
551:
552:
553:
554:
555:
556:
557:
558:
559:
560:
561:
562:
563:
564:
565:
566:
567:
568:
569:
570:
571:
572:
573:
574:
575:
576:
577:
578:
579:
580:
581:
582:
583:
584:
585:
586:
587:
588:
589:
590:
591:
592:
593:
594:
595:
596:
597:
598:
599:
600:
601:
602:
603:
604:
605:
606:
607:
608:
609:
610:
611:
612:
613:
614:
615:
616:
617:
618:
619:
620:
621:
622:
623:
624:
625:
626:
627:
628:
629:
630:
631:
632:
633:
634:
635:
636:
637:
638:
639:
640:
641:
642:
643:
644:
645:
646:
647:
648:
649:
650:
651:
652:
653:
654:
655:
656:
657:
658:
659:
660:
661:
662:
663:
664:
665:
666:
667:
668:
669:
670:
671:
672:
673:
674:
675:
676:
677:
678:
679:
680:
681:
682:
683:
684:
685:
686:
687:
688:
689:
690:
691:
692:
693:
694:
695:
696:
697:
698:
699:
700:
701:
702:
703:
704:
705:
706:
707:
708:
709:
710:
711:
712:
713:
714:
715:
716:
717:
718:
719:
720:
721:
722:
723:
724:
725:
726:
727:
728:
729:
730:
731:
732:
733:
734:
735:
736:
737:
738:
739:
740:
741:
742:
743:
744:
745:
746:
747:
748:
749:
750:
751:
752:
753:
754:
755:
756:
757:
758:
759:
760:
761:
762:
763:
764:
765:
766:
767:
768:
769:
770:
771:
772:
773:
774:
775:
776:
777:
778:
779:
780:
781:
782:
783:
784:
785:
786:
787:
788:
789:
790:
791:
792:
793:
794:
795:
796:
797:
798:
79
```

## handling missing value

13/2

File Edit View Help

1733729216.pdf (15) WhatsApp GitHub Data Har

data/data%20set%20nextbike%20Fdata%20Harmonization\_2.ipynb

Harmonization\_2 Last Checkpoint: 1 hour ago

Kernel Settings Help

Code

JupyterLab

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
7LINES
_and_2[column] = merged_1_and_2[column].clip(lower=lower_bound, upper=upper_bound)

liers handled successfully.")

died 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
```

Search

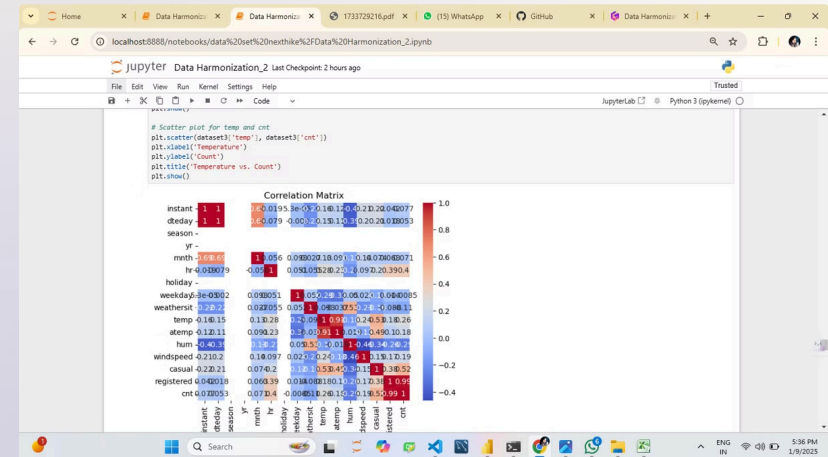
handle outliers

# Compute correlations between attributes to identify relationships.



using matplotlib

using seaborn

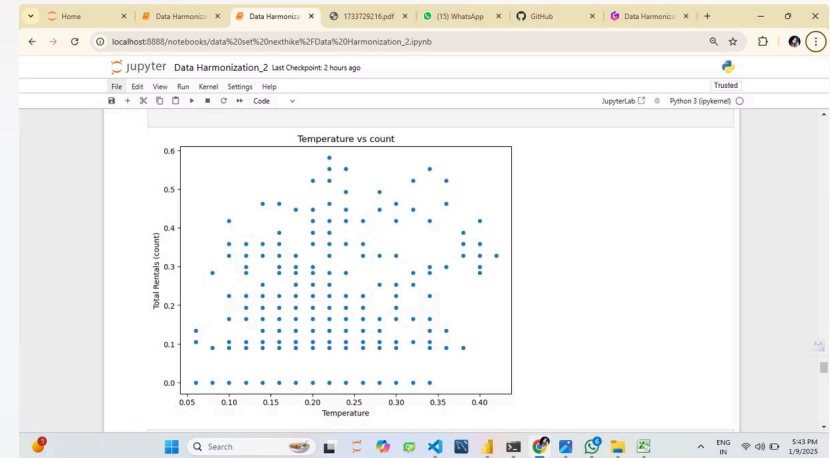


# scatter plots

Visualizations

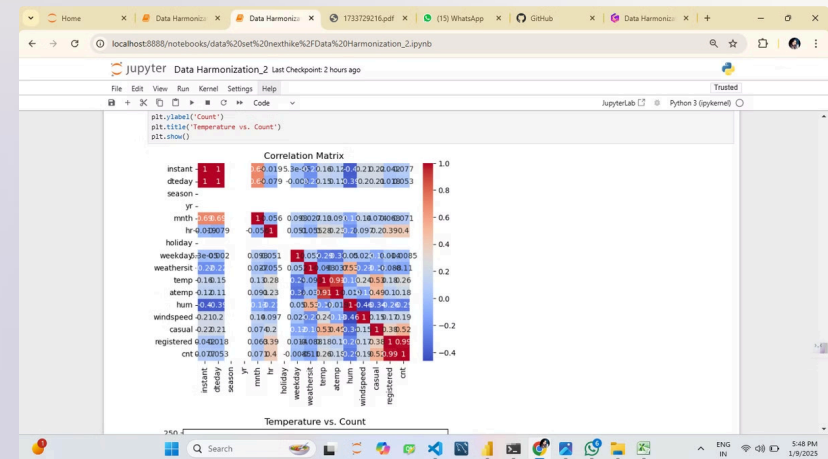
## Scatterplot of Temperature vs Count

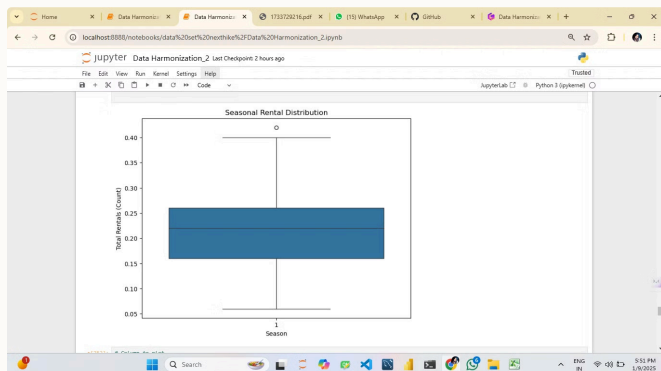
```
plt.figure(figsize=(8, 6)) sns.scatterplot(dataset3, x='temp', y='windspeed')  
plt.title("Temperature vs count") plt.xlabel("Temperature") plt.ylabel("Total  
Rentals (count)") plt.show()
```



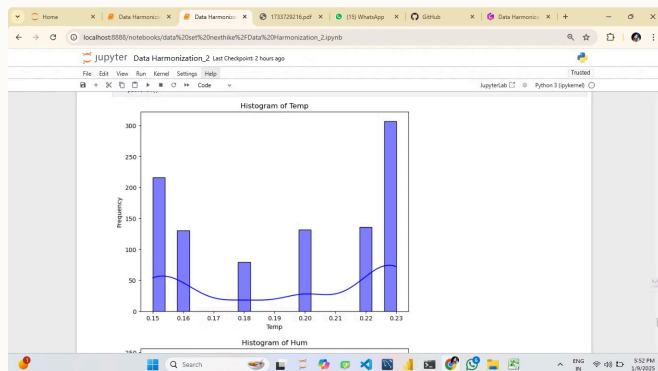
# heatmap

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

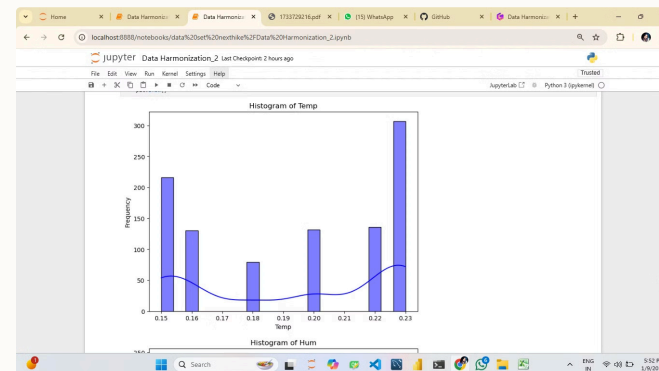




box plot



histogram



Histogram



*Thank You*