# **Data cleaning report for Cyclistic cycling data**

The data had 12 .csv files in total. They each had the data for each month of the year. Below is a detailed report of the data cleaning processes performed on each file individually.

## **202101-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 8,625 empty spaces

Start\_station\_id -> 8,625 empty spaces

End\_station\_name -> 10,277 empty spaces

End\_station\_id -> 10,277 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 103 empty spaces

End\_lng -> 103 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Two entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The two entries were:

-FC1EFEF4475D7F56

-B1235D38EB2F8A9E

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202102-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 4,046 empty spaces

Start\_station\_id -> 4,046 empty spaces

End\_station\_name -> 5,358 empty spaces

End\_station\_id -> 5,358 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 214 empty spaces

End\_lng -> 214 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect and inconsistent values manually, using filter and sort. Checked for negative time using filter and sort. These were no occurrences of incorrect and inconsistent data. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202103-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 14,848 empty spaces

Start\_station\_id -> 14,848 empty spaces

End\_station\_name -> 16,727 empty spaces

End\_station\_id -> 16,727 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 167 empty spaces

End\_lng -> 167 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Two entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The two entries were:

- 5D2797AFFA71B49

- F99335E3A77A57B5

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202104-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 26,056 empty spaces

Start\_station\_id -> 26,056 empty spaces

End\_station\_name -> 16,727 empty spaces

End\_station\_id -> 16,727 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 167 empty spaces

End\_lng -> 167 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Five entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The five entries were:

- BC53ECCBC76278FD

- 209C097828F9CD43

- 6E81034B446FC2FD

- 318DD838369AEA61

- 8ADD13BD8F6A7567

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202105-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 53,744 empty spaces

Start\_station\_id -> 53,744 empty spaces

End\_station\_name -> 58,194 empty spaces

End\_station\_id -> 58,194 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 452 empty spaces

End\_lng -> 452 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Two entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The two entries were:

- 3EC1B5A4D4B9AB99

- 518535DDFA372694

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202106-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 80,093 empty spaces

Start\_station\_id -> 80,093 empty spaces

End\_station\_name -> 86,387 empty spaces

End\_station\_id -> 86,387 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 717 empty spaces

End\_lng -> 717 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Five entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The five entries were:

- 732D64DAD2CC9B73

- 24CAFC421D642C22

- 4E88151C0FBCA967

- 126E4DA0FA0A3E11

- A18D39992AA99793

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202107-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 87,263 empty spaces

Start\_station\_id -> 87,263 empty spaces

End\_station\_name -> 93,158 empty spaces

End\_station\_id -> 93,158 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 731 empty spaces

End\_lng -> 731 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Twelve entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The twelve entries were:

- E6CAF61273BC824D

- 18FA8EAE88922DAF

- 3C634446B054AB44

- D22A7B1DAC07DBEC

- CBBDF121760541A8

- 9180A7126B192284

- AC7176564B0C716F

- B30086B9F4A49AA93

- 2C8922585E5F5B76

- C74A0859A1737B0F

- 2AB142C32C184032

-FEA10221DC06E355

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202108-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 80,093 empty spaces

Start\_station\_id -> 80,093 empty spaces

End\_station\_name -> 86,387 empty spaces

End\_station\_id -> 86,387 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 717 empty spaces

End\_lng -> 717 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Twenty nine entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The twenty nine entries were:

|  |
| --- |
| - 2DBE0C0012129B1 |
| - 0291C81C6932E735 |
| - 4DD89B8847CF6040 |
| - ACD13D8A68452042 |
| - 2F600E738BE2F08A |
| - F33F5D42F4418061 |
| - E377D242052A190D |
| - 42CCBA4971E256F4 |
| - 8DE914BA31FDA368 |
| - 9ACB997E90B54988 |
| - B2A56072F18F4FD9 |
| - 440C3F52F65C5BCD |
| - 84DE853DC2C25C56 |
| - 6F5F7FD4D13ECEA0 |
| - 8610961166DA8412 |
| - 099A7214A2728ACC |
| - 121C9511AE8C11CA |
| - 4E8752F0A62BA2A1 |
| - 5B4D4A00D6A4DAAC |
| - 6DCF2C6EC660CAF9 |
| - 641EF17482044F18 |
| - FB0519FF5907CACC |
| - 23D6456F1268AF00 |
| - E64AE57BDD6FE595 |
| - 9200647C37FD38B0 |
| - 4F67BFFB680CEDCD |
| - D35A56D5648A785A |
| - 9121D4C66D81CDA2 |
| - 3CFA7D4B92C4FDD6 |

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202109-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 93,113 empty spaces

Start\_station\_id -> 93,113 empty spaces

End\_station\_name -> 99,261 empty spaces

End\_station\_id -> 99,261 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 595 empty spaces

End\_lng -> 595 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Thirty six had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The thirty six entries were:

- 88BA6AC7326B179B

- 7C194D321B885287

- 082A465904C062EF

- FA44EBC79CC39AEF

- D4A976387C9C88CE

- DB9D07704E3285A8

- 06576F6E68A4B2C9

- 3AEF4FE696696C8F

- FA4DC99A39C36D54

- 6E5FD2F624AC87D3

- 7F7679196D0D3540

- 393F1F0F9D2D5E21

- 1C374131EFC52365

- BE93718DC9182ED6

- 3A47A9240BC6727E

- 85BC495341AB2F18

- D5E5F806D0FFBC1C

- 41725EA31D3BB35A

- CAE29510F94F312A

- 137BEA38C2312200

- 2EDC266B1A809E5E

- 5F9FAEC899BAFDCF

- A6CF3312091615EE

- F135C632A459FA86

- 31B5BE372F009A2B

- 008CF56DDEA16F7E

- CD7BE84F6552E7AB

- 7623D85B84ADCC27

- 3E68DE63824DF361

- 448610692B74D6B7

- EBD402BB8BF12337

- 6ACD4446FA7A9C6F

- 4EB4634DA7EC8CC7

- CA0E5960C9FC5B28

- 4A68473D329D45C9

- 35D0A35E9454B0F5

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates

## **202110-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 108,210 empty spaces

Start\_station\_id -> 108,210 empty spaces

End\_station\_name -> 114,834 empty spaces

End\_station\_id -> 114,834 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 484 empty spaces

End\_lng -> 484 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect and inconsistent values manually, using filter and sort. Checked for negative time using filter and sort. These were no occurrences of incorrect and inconsistent data. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202111-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 75,290 empty spaces

Start\_station\_id -> 75,290 empty spaces

End\_station\_name -> 79,187 empty spaces

End\_station\_id -> 79,187 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 191 empty spaces

End\_lng -> 191 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect values manually, using filter and sort.

- Checked for negative time using filter and sort. The rows I found are listed below.

**Negative amounts of time:**

Fifty three entries had negative time in this file. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon. The fifty-three entries were:

- 7CA158F5F050156E

- FD8AF7324ABAE9DA

- 508B09A5FB0737DC

- 6F9E76F5EDAAC1B8

- 7AECC76D1562B51C

- B506DCD44974C575

- CDB307B8494885AD

- FFD5A2DDE1FAAA90

- 7E24361D78747AF6

- 53222CFE6657D53D

- 5AA2BC364BC7A569

- 6A2DCA5CB1596CA6

- 9D8C6B064E2ECEDE

- 46BC473D0F68B4A2

- 83B6C6B10248241E

- D6D3692804119896

- 3556173EA13B5C84

- 2B22D0C575CC2DF8

- 08F7F73C0EC09A0B

- 46D20A622D89BDCE

- 82ACDD5EE3627B6C

- D631251FA9C7FC03

- 49DF29FDEC7E48FB

- 984A42435A4BBEB0

- E17513A83D75F889

- 9BEB5FBD0EF2FCFD

- 5132D82DCD66BA95

- 0C29F2FD22BFE9C3

- 83778DB379E8B04B

- 36BCBD47831CECD3

- 36198116E86D6457

- E16920EC561FFEBD

- 021DC77C70A3E367

- EDBA8663286116BE

- B60E253706E151AF

- 658F94C7AE8F182C

- CF4D877FEB218570

- 1E08692D5C77F7AA

- 287605D0D79731E6

- B029250A1EFF2975

- 802BFC5FF150F1F8

- F4E4485BFB33D916

- E394AA8418D47B74

- F2AF011E90AC8918

- 0006AA1377F8BCA1

- E89DD4EBFBD231E3

- 8598DCE30427F984

- 600852E2195647B8

- 1374B0438900E9E9

- 214464733718183D

- 36CAA82BB5852EF4

- D2220642A2BF783C

- 235ACD294AFB837F

These were the only occurrences of incorrect and inconsistent data, I deleted them after location and recording them.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

## **202112-divvy-tripdata**

**Empty data:**

Start\_station\_name -> 51,063 empty spaces

Start\_station\_id -> 51,063 empty spaces

End\_station\_name -> 53,498 empty spaces

End\_station\_id -> 53,498 empty spaces

All of the above empty spaces were filled with the word “unknown”.

End\_lat -> 144 empty spaces

End\_lng -> 144 empty spaces

All of the above empty spaces were filled with the word “null”.

- Checked for incorrect and inconsistent values manually, using filter and sort. Checked for negative time using filter and sort. These were no occurrences of incorrect and inconsistent data. A lot had a zero amount of time but I elected to keep them in the file since this could mean the person elected not to ride the bike but still paid thus it is a normal phenomenon.

-There were no duplicate rows recorded. I checked for duplicates using conditional formatting and filters

-There were no outdated data entries in the file. All the data was within the month of the file. I checked using filters on the start and end dates.

|  |
| --- |
| Citation: Divvy Bikes, divvy-tripdata, Divvy Bikes. |