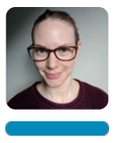


Data Cleaning Challenge: Handling missing values

Python notebook using data from multiple data sources · 78,787 views · dailychallenge, multiple data sources

Data

Data Sources

- Detailed NFL Play-by-...
  - NFL... 362k x 102
  - NF... 408k x 102
  - NF... 449k x 255
- San Francisco Buildin...
  - Build... 199k x 43
  - DataDict... 43 x 3

Detailed NFL Play-by-Play Data 2009-2018

nflscrapR generated NFL dataset wiith expected points and win probability

Last Updated: a month ago (Version 6)

About this Dataset

Introduction

The lack of publicly available National Football League (NFL) data sources has been a major obstacle in the creation of modern, reproducible research in football analytics. While clean play-by-play data is available via open-source software packages in other sports (e.g. nhlscrapr for hockey; PitchF/x data in baseball; the Basketball Reference for basketball), the equivalent datasets are not freely available for researchers interested in the statistical analysis of the NFL. To solve this issue, a group of Carnegie Mellon University statistical researchers including Maksim Horowitz, Ron Yurko, and Sam Ventura, built and released nflscrapR an R package which uses an API maintained by the NFL to scrape, clean, parse, and output clean datasets at the individual play, player, game, and season levels. Using the data outputted by the package, the trio went on to develop reproducible methods for building expected point and win probability models for the NFL. The outputs of these models are included in this dataset and can be accessed using the nflscrapR package.

Content

The dataset made available on Kaggle contains all the regular season plays from the 2009-2016 NFL seasons. The dataset has 356,768 rows and 100 columns. Each play is broken down into great detail containing information on: game situation, players involved, results, and advanced metrics such as expected point and win probability values. Detailed information about the

Run Info

Succeeded	True	Run Time	24.7 seconds
Exit Code	0	Queue Time	0 seconds
Docker Image Name	kaggle/python(Dockerfile)	Output Size	0
Timeout Exceeded	False	Used All Space	False
Failure Message			

Log

Download Log

```
Time   Line #  Log Message
3.7s        1  [NbConvertApp] Converting notebook script.ipynb to html
3.7s        2  [NbConvertApp] Executing notebook with kernel: python3
23.9s       3  [NbConvertApp] Writing 294163 bytes to __results__.html
23.9s       4
23.9s       6  Complete. Exited with code 0.
```

Comments (317)

Filter/sort ›

Please sign in to leave a comment.

Wei Chen • 6 months ago

1

Very useful, thank you!

Options

perkleeee • 6 months ago

1

Thank you for great tutorial !

Options

Cyringa • 6 months ago

1

Very useful challenge for novice. Thank you!

Options

VSS • 6 months ago

1

Completed Day 1 Challenge.
Thankyou Rachael . It is really helpful and informative :) .

Options

Joshua • 6 months ago

1

Rachel, thanks for putting this challenge together! It is perfect for beginners like myself.

Options

shreya • 7 months ago

3

Hi Rachael! I think that was a wonderful introduction to data cleaning.
Could you please explain the below syntax and its important uses?
np.random.seed(0)
Thanks!

Options

Rachael Tatman • 7 months ago 3

JamesZhou • 7 months ago 1

THX ! :)

Options

Marina • 7 months ago 1

thanks!

Options

tecnoholic • 7 months ago 1

Good tutorial Rachel. Thank you.

Options

Imran Iskandar • 7 months ago 1

Thank you! Waiting for tommorow's challange!

Options

Jake Nocentino • 7 months ago 1

Thanks for this tutorial!

Options

Artem Prosvirnin • 7 months ago 1

Thank you Rachael. The tutorial was very useful.

Options

Lim Chia Hooi • 7 months ago 1

Thanks for this tutorial!

Options

Mandy • 7 months ago 1

Thanks Rachael!
Here is my work: https://www.kaggle.com/tcmandyw/data-cleaning-challenge-handling-missing-values
Keep cracking down other challenges^^

This line set the random seed for numpy.

This answer that when

Options

Parichita Nandi • 7 months ago 1

Thank you Rachael. The tutorial was very helpful.

Options

Mohd Akhtar • 7 months ago 1

Thanks, Was my first experience with python, went down very smoothly

Options

shreya • 7 months ago 1

Thank-you Rachel for such a good tutorial.

Options

zhouleyuan • 7 months ago 1

Great! This is really helpful for beginners like me, thanks a lot. I have two questions. First, how can I format the output table with python3 when there are too many columns? Like with first 10 columns and last 10 columns displaying in detail, and other columns displaying with ellipsis. Second, it's about filling missing values automatically. Is it recommended that we replace missing values with value coming directly after it in the same column? How about with value before it in the same column? And how to cope with the missing value in the bottom or top of the column?

Options

Rachael Tatman • 7 months ago 0

Tarak • 8 months ago 1

This is so good tutorial, thanks Rachael Tatman
Looking forward to more challenges.....

Options

shonafeng41 • 8 months ago 1

This challenges tutorial is super helpful!

Options

Mahbub • 8 months ago 1

Excellent!

Options

Berk Atik • 9 months ago 1

Your 5-day challenges are awesome for the beginners like me, it helps to wrap my head around the concept. Thank you a lot :)

Options

Akash Agarwal • 9 months ago 1

Thanks a lot! this was very helpful.

Options

DavinderKumar • 10 months ago 3

Great learning ....waiting eagerly for tomorrow's session :)

Options

Prableen Kaur • 10 months ago 4

Nice tutorial Rachel!
I like how you are trying to get us to build an intuition of how to do "imputation".
Looking forward to tomorrow's lesson!
Here's my solution: https://www.kaggle.com/prableen20/data-cleaning-challenge-handling-missing-values

Options

Hannah Catrin • 10 months ago 1

Also, I hope to participate in more challenges in future :)

Options

Hannah Catrin • 10 months ago 1

Thanks for creating the 5 day challenge, I learned a lot!

Options

Gopesh Dwivedi • 10 months ago 1

This was a great tutorial, thanks :)

Options

pawan bansal • 10 months ago 1

https://github.com/hackbansu/kaggle-5-day-data-cleaning-challenge

All my notebooks link.

Thanks Rachael for this great challenge.

Options

gcoolt • 10 months ago

1

This was very helpful. Thanks

Options

Zayed Shah • 10 months ago

1

Thanks Rachael

Options

Praxitelis-Nikolaos Kouroupetroglou • 10 months ago

1

thank you Rachael for the data-cleaning challenge day-1, capital work!

Options

Vibha • 10 months ago

1

Really good introductory tutorial Rachel! Having done these basics and watching them execute is a tiny first step in my Data Science journey. Thank you :)

Options

Options

vica • 10 months ago

1

Nice work thank you.

Options

Arvind Kumar • 10 months ago

1

This Topic really makes a sense at Kaggle...

Options

Rich King • 10 months ago

1

Thanks for this tutorial. It's super helpful!

Options

Ryan Juliano • 10 months ago

1

to echo many other comments, this was a nice tutorial. Thanks for investing the time to make it. I look forward to completing the subsequent parts (which I've sadly fallen a bit behind on!)

Options

Aman Saxena • 10 months ago 1

I'm not able to find an R notebook for this, can anyone help me out? @Rachael, you've only uploaded the Python notebook, right ?

Options

Rachael Tatman • 10 months ago 1

Shehjar Kaul • 10 months ago 1

Great tutorial to start with! Here is my code!

Options

Vipin Kumar • 10 months ago 1

Thank you for the great tutorial, and providing the suggestion for further learning from Dan's notebook.

Options

jaludac • 10 months ago 1

It’s great!
Thank you Rachael

Options

konichuvak • 10 months ago 1

Handling NaN's is crucial! Thanks for doing the series!

Options

Samarendra P • 10 months ago 1

Thanks Rachel !! The content was great and the datasets were good !!

Options

Rahul Singh • 10 months ago 1

Is anyone trying these out in R?

Options

Rachael Tatman • 10 months ago 0

RohithB • 10 months ago 1

Thanks for the great tutorial @Rachael. Please check my Day1 solution : Handling missing values

Options

rory81 • 10 months ago 1

Thanks Rachel! Not only am I learning a lot about data cleaning, but as a Python newbie also a lot about using Python.
So here is my Solution

Options

Masum Rumi • 10 months ago 1

Thank you so much @Rachel. Here is my Solution. Doing day#2 now and looking forward to the rest of them.

Options

John Olayisade • 7 months ago 0

Options

Dany Venero • 10 months ago 1

Learning a lot...Thanks Rachel!

Options

Yong • 10 months ago 1

thanks

Options

Dare • 10 months ago 1

Nice!

Options

Hernan Goldenberg • 10 months ago 1

Thanks a lot Rachel, I'm taking my first steps in data science.....
https://www.kaggle.com/ruso72/data-cleaning-challenge-handling-missing-v-ea4b91

Options

Banjo • 10 months ago 1

Thanks Rachel! I really enjoyed this first part and hope other parts will be as simple as you did to this first part. Your turn solution link

Options

ColaCole • 10 months ago 1

My Teacher thank you very much for your education Ms. Rachael Tatman. Great work.

Options

Said Aspen • 10 months ago 1

Good stuff, thanks!

Options

Sabu Joseph • 10 months ago 1

Hi Rachea l,
Updated the notebook with analysis around Street Number Suffix and Zipcode. Analysis around zipcode can be productive since it may be connected with missing Location or inherent issues in the source system that creates sf_permits dataset while collecting these tow fields or if zipcodes are extracted based on Location using wrapper applications. Another assumption is the growing importance of Location coordinates in many applications

Will post finding from datasets about San Fransisco on the Datasets listing

Options

hariprasaath • 10 months ago 1

hope it helps. https://www.kaggle.com/hariprasaath/data-cleaning-challenge-handling-missing

Options

Bellamy • 10 months ago 1

This is great, thanks for this tutorial. I'm looking forward to the next few days.

Options

Jessica Chopra • 10 months ago 1

Great Tutorial! I have used tools like Angoss and Enterprise Miner to impute missing values. These tools can create decision trees based upon certain conditions and we can impute missing values based upon that data in the neighborhood records.

Options

WangYong • 10 months ago 1

Thanks for sharing, it is a great kickoff.

Options

Here is my Kernel

Features:

1. Visualization for missing data
2. Ignorable & Non-ignorable concept

Options

Semloh • 10 months ago

1

Thank you Rachael was really a great tutorial. Looking forward for the next lesson.

Options

naveensaliyan • 10 months ago

1

Thank you Rachel, good refresher on data cleaning, waiting for advanced cleaning techniques in next tutorials.
I never knew there is a **sample** method for DF, thanks for introducing :)

Options

Deep Thought • 10 months ago

1

Great tutorial. Thanks Rachel!

Options

GSD • 10 months ago

1

Great tutorial Rachel..Thanks

Options

Elvirasun28 • 10 months ago

1

Pretty interesting and useful to a new fresher. Please check my solution
https://www.kaggle.com/elvirasun28/data-cleaning-challenge-handling-missing-values

Options

Punyajoysaha • 10 months ago

1

nice way to learn data cleaning

Options

Anthony Wu • 10 months ago

1

Very helpful for beginners like me. Thanks!

Options

LuizGermano • 10 months ago

1

Thanks Rachel!!! It´s an amazing tutorial. I´ll try the lessons!

Options

Javed Sheikh • 10 months ago

1

thanks for making this tutorial (like the rows we read at the start)

Completed!

Options

M.G.Hunter 10 months ago 1

I'm still new to data science and have little knowledge of coding. Yet, this was easy to follow and understand. Thanks Rachael

Options

Manojna Kalapala • 10 months ago 1

Needed this challenge!!!

Options

Peiqin • 10 months ago 1

Thanks Rachael for the nice tutorial! My solution: https://www.kaggle.com/peiqin/data-cleaning-challenge-handling-missing-values

Options

Arash Koupaei • 10 months ago 1

Thank you Rachael - interesting competition - my solution to the challenge

Options

KotsirasDimitris • 10 months ago 1

Nice tutorial!thank you! https://www.kaggle.com/jimintel/data-cleaning-challenge-handling-missing-values

Options

Imron1 • 10 months ago 1

This is very interesting. Thanks Rachel.
this is my solution:
https://www.kaggle.com/imron1/data-cleaning-challenge-handling-missing-v-4ba114
feel free to have any feedback.

Options

Ace Palmero • 10 months ago 1

Great Tutorial! I learned a lot. A quick question about imputation. Doesn't sucha broad fix add a lot of incorrect infromation to your dataset. For example, with the street number suffix, a lot my values in that column are filled as 'A'. I assume the street number suffixes should be different so isn't it preferable to have Nan instead of incorrect values.

Options

i
n
o
r
d
e
r
)
W
h
e
t
h
e
r
t
h
e
v
a
l
u
e
b
e
f
o
r
e
o
r
a
f
t
e
r
m
a
k
e
s
m
y
s
e
n
s
e
,
o
r
i
f
y
o

Rachael Tatman • 10 months ago 0

sdnagaraj • 10 months ago 1

Instead of 'bfill', isn't it better to fill based on similarity to a different column value?

Options

Rachael Tatman • 10 months ago 4

Andre Becklas • 10 months ago 1

Very nice! Thank you for your time doing this, Rachael :)

Options

SmitM • 10 months ago 1

Fun way to learn python data cleaning! Looking forward to the rest of the week. @Rachael - Awesome work!

Options

Dipika Baad • 10 months ago 1

It was a nice challenge. Thank you Rachael. Looking forward for tomorrow's challenge ! Here is my notebook

Options

Dimitri Saingre • 10 months ago 1

Thanks ! Pretty interesting and well explained :)

Options

Rozzah • 10 months ago 1

Great work @Rachael. Looking forward to challenge 2.

Options

Anezka Kolaceke • 10 months ago 1

Thank you for the tutorial! My solution: Handling missing values

Options

Anezka Kolaceke • 10 months ago 1

Thank you for the tutorial! My solutions: Handling missing values

Options

Saranjit saikia • 10 months ago 1

nice tutorial learned a lot

Options

Hugh Zabriskie • 10 months ago 1

Thanks for the tutorial, Rachel! It seems that the NFL data doesn't have the column descriptions in the documentation as mentioned. The link to the data in the tutorial is: https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016. I went to Data > Column Metadata, but didn't see the column descriptions.

Options Options

Rachael Tatman • 10 months ago 0

uday • 10 months ago 1

An excellent tutorial looking forward to next lesson. Thank you, Rachael, for this series. My solution: https://www.kaggle.com/udaykiran3207/data-cleaning-challenge-handling-missing-v-8af847

Options

Prashant Brahmbhatt • 10 months ago 1

As a rookie this was really really helpful. Thanks!

Options

ameer.khosoof • 10 months ago 1

Excellent job.. a very helpful tutorial for beginner.

Options

snehanshusengupta • 10 months ago 1

Nice article

Options

Prashant Kishore • 10 months ago 1

NIce tutorial. Thank you!
https://www.kaggle.com/prashant001/data-cleaning-challenge-handling-missing-values

Options

Anand Panchal • 10 months ago 1

Thank you for the wonderful tutorial Rachael. I am looking forward to the next challenge. Here's my notebook for day 1: https://www.kaggle.com/apanchal2/data-cleaning-challenge-handling-missing-values

Options

Options

Jordan • 10 months ago 1

edit -solved

Options

Atpaul • 10 months ago 1

Thank you Rachael. Looking forward for Day 2 of the Challenge

Options

MonalisaMishra • 10 months ago 1

Really useful in refreshing the concepts. Thanks, Rachael!!
https://www.kaggle.com/mmishra90/data-cleaning-challenge-handling-missing-values

Options

huangkx • 10 months ago 1

It is easy to have a handle on data cleaning by following the notebook, thanks!!

Options

TheRunningPhysicist • 10 months ago 1

As a new guy on Kaggle I find this kind of competiotions/lessons to be really useful. I really appreciate the work and commitment in this 5-days challenge!

If somebody wants to take a look at my notebook you are more than welcome: https://www.kaggle.com/therunningphysicist/data-cleaning-challenge-handling-missing-values/notebook

Can't wait for next lesson!

Options

Sidharth • 10 months ago 1

Great work !!
I joined Kaggle recently and was skeptical about signing up for this tutorial, but really enjoyed and learnt a lot.
Looking forward to upcoming tutorials.

Glad to share my work : https://www.kaggle.com/sidharthsuman/data-cleaning-challenge-handling-missing-values

Options

Ibrxk • 10 months ago 1

Great Tutorial here is my solution :-) https://www.kaggle.com/ibrahimkochbati/data-cleaning-challenge-handling-missing-values

Options

Steven Hu • 10 months ago 1

Thank you, Rachael!! This hands-on practice is super helpful for beginners, especially people who have a dream of data scientist. Thank you! Cannot wait for day two challenge! here is my notebook for day1

Options

Anthony Jatobá • 10 months ago 1

Nice tutorial, Rachael! The examples were fine and my poor python wasn't a obstacle.

Here's my notebook: https://www.kaggle.com/anthonyjatoba/data-cleaning-challenge-handling-missing-values

Options

Jonas Conde • 10 months ago 1

Nice tutorial!

Is there any way to import the csv files from the OpenAddresses kernel?

Options

Rachael Tatman • 10 months ago 1

Saroj • 10 months ago 1

Thanks Racheal
https://www.kaggle.com/saroj13278/data-cleaning-challenge-handling-missing-values

Options

Patrick Kakou • 10 months ago 1

Nice challenge, I am just for Day 1 and ready for tomorrow. Thanks Rachael

Options

AhmedSultan • 10 months ago 1

Thanks for the tutorial! ... waiting for tomorrow's tutorial

Options

the second finana(0) makes any value to add the data as missing after the first

Wei Chun Chang • 10 months ago 1

Hi Rachael, thanks for the tutorial. I love your challenge :)

Here are my fork notebook and some review. Welcome for any suggestion and comment.

https://www.kaggle.com/justjun0321/data-cleaning-challenge-day-1-done-review

Options

SarahXu • 10 months ago 1

Thank you ! :)

Options

Sabu Joseph • 10 months ago 1

Great experience..thanks for explaining the concepts in deep.
How can I access Iowa housing data to try with methods DanB suggested ? i tried with path pd.read*csv('../input/(iowa-housing-snapshot/iowa*data.csv') but getting filenot found error

Options

Rachael Tatman • 10 months ago 1

Demongolem • 10 months ago 1

I think maybe I am not so familiar with notebooks. When I click in the first cell that I see that loads all the data, I get a NameError: name 'np' is not defined error. It says [2] to the left, so is there a [1] somewhere that I am missing?

EDIT It works now, but for me it seems I have to run several cells multiple times for the correct action to be taken. For example, I just printed out the pct of missing values for the sf_permits data set, but it would not spit out the correct number until I ran it 3 times.

Options

FSBDS_AndreMarquesLeite • 10 months ago 1

Thanks Rachael, Very detailed and clear. I'm also looking fwd for tomorrow´s lesson ;)

Options

GranVisir • 10 months ago 1

Thank you Rachael! Very Socratic lesson to learn: "Know your data"!

Happy to share my work: Cleaning Data Challange: Day 1

Options

John Xu • 10 months ago 1

Very enjoyable study. Thanks. Here is my notebook John's Notebook

Options

```
g = 0
t
h )
e :
f
i
l
l
n
a
(
0
)
```

Sergey Proshuta • 10 months ago

1

Great tutorial and cool start to the first day!
I tried to add some extra things to my solution:
https://www.kaggle.com/proshuta/data-cleaning-challenge-handling-missing-extra/
Cannot wait for the next day :)

Options

Tanveer Baba • 10 months ago

1

It was great. Honestly, After first lesson i'm excited for tomorrow's lesson
Thank you Rachael
Thank you very much

Options

Shridhar • 10 months ago

1

Great learning, thank you.
Regards,
Shridhar

Options

TFu • 10 months ago

1

I really like the way reasoning behind the missing value, I usually just remove the rows or columns before this tutorial

Options

SK • 10 months ago

1

Thanks for the detailed explanations. Looking fwd to tomorrow's session

Options

AimeShangula • 10 months ago

1

thanks for t(o)uts Rachel

Options

manasa • 10 months ago

1

Great Tutorial!! Thanks Rachael. Looking forward for more.

Options

Abdelrahman Hamza • 10 months ago 1

Thank you! amazing day :) Looking forward to tomorrow's lesson.

Options

Xenophontas Psichis • 10 months ago 1

Really nice and helpful tutorial!

Options

dilex • 10 months ago 1

Thanks! Very good first steps tutorial.

On a side note : there are missing "s" in sf_permits in the end of paragraphs 2 &5 ;)

Options

Rachael Tatman • 10 months ago 0

Thank you for pointing out the missing "s"!

LennartGrosser • 10 months ago 1

Cool notebook on an important topic which can really make a difference in terms of accuracy when done well.

I'd further suggest looking at the distribution of a column when replacing missing values.

Basic example: if the values of a column are equally distributed you want to impute differently than if they were normally distributed.

For categorical values one can look at the histogram.

Options

Rachael Tatman • 10 months ago 1

Andreas Paul • 10 months ago 1

Great lesson, thanks a lot! Looking forward to more!

Options

Marco Rigo • 10 months ago 1

Nice tutorial!

very clear

Options

Hui Miao • 10 months ago 1

Thank you Racheal! Great lesson for someone who is new to data cleaning!

Options

Options

Eladio Rego • 10 months ago 0 1

Great!!! I do like it!

Options

Rohit Singh Adhikari • 10 months ago 1

Thanks for the explanation....will follow the coming sessions!!!

Options

Sunny Chidural • 10 months ago 1

Great.. :)

Options

Chayan Shrang Raj • 10 months ago 1

Awesome Challenge! Simple and effective data cleaning techniques for beginners. Happy to share my notebook https://www.kaggle.com/chayanraj/data-cleaning-challenge-handling-missing-values.
Never Stop Learning...

Options

LakshyaRoy • 10 months ago 1

Great tutorial

Options

Arunava • 10 months ago 1

It was a nice walkthrough. Thanks for this :)
Here's the link to my fork for Day 1 of Data Cleaning Challenge

Options

ChrisBow • 10 months ago 1

Thanks Rachael, I'd been looking a bit of a more structured way to work on my Python (not that I'll forget about you, R!), and this was spot on! Looking forward to the rest of the week.

Here's my notebook for day 1 (I hope I imputed the missing value at the end of one of the questions correctly!)

Options

Ibio • 10 months ago 1

useful skills

Sabu Joseph • 10 months ago 1

Thanks for the explanation....will follow the coming sessions!!!

Options

Options

CRK • 10 months ago

1

Great tutorial Rachel. Thank you. Excited to move on to Day 2!
This course reminded me of an interesting project that i was working in 2011 where we had customers in Dublin,Ireland. I will not go in detail about the project but part of it was to deliver points cards to the customers address.
I did not knew until then that Dublin,Ireland did not have any postcode... the records in the database showed 'NaN' for few postal codes and few others had 'text and numbers' mixed together. I was trying to figure out for a long time if it was "does not exist" or "wasnt recorded".
But later in 2014,ireland started to introduce the postal code.
https://qz.com/272332/ireland-is-just-now-getting-around-to-introducing-postal-codes/

Options

Options

Rachael Tatman • 10 months ago

0

Sabu Joseph • 10 months ago

0

Feiqi Zhang • 10 months ago

1

nice teaching style that is well organized and super easy to do and follow...

Options

Somrik Banerjee • 10 months ago

1

Thank you Rachel for pointing out the subtle difference between data not being recorded and nonexistence. I am familiar with most of this content otherwise. Looking forward to the next session.

https://www.kaggle.com/somrikbanerjee/data-cleaning-challenge-handling-missing-values

Options

NehaS • 10 months ago

1

Was amazing lesson... thanks a lot. Looking forward to more lessons.

Options

Moni Gill • 10 months ago

1

Nicely commented and explained. Looking fwd to other sessions.

Options

时间刺客 • 10 months ago

1

I'm a novice.我觉得这是很好的学习教程。没有什么练习的机会，刚好可以试试手。很期待明天的挑战！！！

Options

nengkuan2 • 10 months ago 1

very good reasoning on why/when/what to drop/impute. But I am still confused about NAN and None.

1. It looks to me that dropna() only remove NAN. It does not remove None. Am I right?
2. Is there a way to remove None?
3. My understanding about NAN is that it is a result from floating operation like 0/0. We are also using NAN to represent missing or empty. Does None also mean missing or empty?

Options

junnu • 10 months ago 3

Niyamat Ullah • 10 months ago 1

Options

Actual learning is doing in practice. And Kaggle is doing the same thing. Thanks Kaggle!

Jonas Conde • 10 months ago 0

Options

Weimin Yu • 10 months ago 1

Tkanks！ ,it's very useful for learner! It's my solution :
https://www.kaggle.com/yuweiming70/data-cleaning-challenge-handling-missing-values

Options

Djaballah • 10 months ago 1

Well-explained tutorial, Looking forward to the next challenge.

Here is my notebook
Data Cleaning Challenge: Handling missing values Day 1

Options

Kelechi • 10 months ago 2

Great job, Rachael. I just published an article about my experience here -
https://medium.com/@Kelechukwu_/a-practical-guide-to-data-cleaning-for-beginners-bd4f45967825

Options

Richard Allen • 10 months ago 2

Thanks for the kernel Rachel, just a quick note - for the 'More Practice!' section, I've found the data for San Francisco addresses to be missing in what seems like a deliberate exclusion. I have a brief analysis here:
https://www.kaggle.com/bigironsphere/missing-san-francisco-data

And the missing data itself can be directly downloaded here:
http://results.openaddresses.io/sources/us/ca/san_francisco

Options

Options

Rachael Tatman • 10 months ago 0

Kristen O Anderson • 10 months ago 2

Everything is so well explained and the steps are so logical. Thanks!

Options

Lewis Tunstall • 10 months ago 2

In addition to calculating the total fraction of missing values in a dataset, I often find it's useful to visualise the percentage of missing values per feature. In case this is useful to others, the following code snippet produces a bar plot for the `sf_permits` dataframe:

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

# get percentage of missing values
df_na = (sf_permits.isnull().sum() / len(sf_permits)) * 100

# drop columns without 1 alues and sort
df_na = df_na.drop(df_na[df_na ==
0].index).sort_values(ascending=False)

# create plot
f, ax = plt.subplots(figsize=(12, 8))
plt.xticks(rotation='90')
sns.barplot(x=df_na.index, y=df_na)
ax.set(title='Percentage of missing data by feature',
ylabel='Percentage missing')
plt.show()
```

Options

ayoi • 10 months ago 0

Chirag Sehra • 10 months ago 2

Great Work Rachel! The doubts and comment section shows what a great teacher you are. Thanks for the efforts to make everyone understand such important topics that people mostly skip.

Options

Made in Russia • 10 months ago 2

nengkuan2 • 10 months ago 1

Thanks for tutorial. I`m kind of late bird, but anyways that was a good start.
Thats my notebook https://www.kaggle.com/madeinrussia/data-cleaning-challenge-handling-missing-values/notebook

Options

Aahan Singh • 10 months ago 2

Thank you for the tutorial Rachael. Here's my solution
https://www.kaggle.com/aahansingh/data-cleaning-challenge-handling-missing-values

Options

jaishiva • 10 months ago 2

Guys we can get the StreetNumberSuffix >> based on description ; and Street. Below are the examples

In the description we can find the Address information based on which we can get the StreeNumber Suffix

based on below concept

A number suffix is a letter that might come after an address if there aren't enough numbers for all the buildings on a street. (For example, if your address is 9A Main Street, the suffix would be "A".)

below is the link you can have a look at
http://www.mississauga.ca/portal/helpfeedback/faq?paf_gear_id=12000021&itemId=104500575n&action=faqAnswer

we can update the data
StreetnumberSuffix / Street
A >> Minna/Dolores/Pearl/Rausch >> Street space >> for blanks

V >> Hampshire / Hallam / 06th >> Street space >> for blanks

C >> Hayes / Montgomery / Stockton

Options

SHAF • 10 months ago 2

In case someone want's to fill the NA's in both direction (forward and backward in chain) can use this code:
subset*nfl*data.fillna(method='ffill',axis=0).fillna(method='bfill',axis=0)

Also, thank you for the tutorial Rachael.

Options

Docty • 10 months ago 2

Hello Rachel, I thank you for your tutorial, However, how we can impute missing categorical values in the data???

Options

Rachael Tatman • 10 months ago 0

Joseph Romani • 10 months ago 2

This is a great tutorial, but I do have to echo Ace Palmero's comments. How you handle missing values in a dataset is entirely dependant on what you are intending to use the data

for and what techniques you are going to apply. Solution is here
https://www.kaggle.com/attackgnome/data-cleaning-challenge-handling-missing-values

Options

Ahmed Alqam • 10 months ago 2

Awesome competition, looking forward to continuing the rest of it.

Options

Gabriel Montañola • 10 months ago 2

For those trying to use OpenAddresses dataset

I'm new to Python (2 months) and Pandas (3 days...I guess..) - so mind any newbie coding

San Francisco is not listed as a City. Don't ask me why:

I googled the range of zipcodes and used this (first part is converting the POSTCODE to numericvalues, then selecting the range with between the values I want to.)

```
ca_addresses['POSTCODE'] = pd.to_numeric(ca_addresses['POSTCODE'],
errors='coerce')

ca_addresses['NUMBER'] = pd.to_numeric(ca_addresses['NUMBER'],
errors='coerce')

new = ca_addresses[ca_addresses['POSTCODE'].between(94102,
94177)].sort_values(by=['POSTCODE'])
```

I found a way to get the Zipcode using the closest number and the street name but I can't get it to work with pandas apply() - mostly because I suck at it.

But it goes like this:

Defining a function to concat the street name

```
def street_concat(row):
    return row['Street Name'].upper() + " " + row['Street
Suffix'].upper()
```

I know that there is some *AV vs AVE* problems, but I can work it later.

Applying this to the dataframe to create a new column with the concat street name

```
sf_permits['COMBO_ST'] = sf_permits.apply(street_concat, axis=1)
```

Def a function to get the nearest number

LennartGrosser • 10 months ago 0

```
def find_nearest(array, value):
    idx = (np.abs(array-value)).idxmin()
    return array[idx]
```

But then I'm stuck. I can get a ZIPCODE value using this:

```
new['POSTCODE'][(new['STREET'] == 'JERROLD AVE') &amp;amp;
(new['NUMBER'] == find_nearest(new[new['STREET'] == 'JERROLD AVE']
['NUMBER'], 2241))].values[0]
```

This returns me **94124**. But I'm hardcoding the *street name* and *number*. I don't know how to build a function to automate this process. Tried this:

```
def zip_finder(row):
    if row['Zipcode'].isna():
        return new['POSTCODE'][(new['STREET'] == row['COMBO_ST'])
&amp;amp;
        (new['NUMBER'] == find_nearest(new[new['STREET'] ==
row['COMBO_ST']]['NUMBER'],
        row['NUMBER']))].values[0]
    else:
        return row['Zipcode']
```

But it fails in so many ways that I'm ashamed to post it here. Someone willing to improve this?

Options

Killdary Aguiar • 10 months ago 2

Very cool, thank you Rachael, I learned a lot. So I have a question, in filling in the data, would not it be better to sort the data through the columns related to the addresses before filling with the dataset's own data?

Here is my solution: https://www.kaggle.com/killdary/data-cleaning-challenge-handling-missing-values

Options

Rachael Tatman • 10 months ago 0

Surekha • 10 months ago 2

Thank you Rachael for this useful tutorial.
I have a question about the fillna() method. If we are replacing all NA's with the value that comes directly after it in the same column, why do we still fill the values with "0"? subset*nfl*data.fillna(method = 'bfill', axis=0).fillna("0"). Is this for the last row, where there are no values after?

Here is my solution: https://www.kaggle.com/surekha09/data-cleaning-challenge-handling-missing-values

Options

Rachael Tatman • 10 months ago 1

Thomas Zoeller • 10 months ago 2

Hi @Rachael, funny and interesting to walk trough. Thanks!
Btw.: On the Mac CTRL+ENTER works for executing the code or SHIFT+ENTER to execute and jump to next section.

Options

Rachael Tatman • 10 months ago 0

Nitin • 10 months ago 0

mememem • 10 months ago 2

"More practice!" section is a must read, because strategies presented in main notebook are not that advanced.

Options

DEBJIT GHOSH • 10 months ago 2

This was my **first kernel** on Kaggle. Not only did I learn about handling missing data in Python, but I also learnt how to write, edit and publish kernels. Great learning I must say. Thank you Rachael for the tutorial! Looking forward to Day 2 of the challenge.
Here's my notebook: https://www.kaggle.com/debjitghosh/data-cleaning-challenge-handling-missing-values

Options

rush • 10 months ago 2

Thanks a lot, Rachael, nice tutorial!

My fork is here.

But OpenAddresses dataset doesn't contain data for San Francisco.

Options

Rachael Tatman • 10 months ago 0

Mihai • 10 months ago 2

I was getting this error when reading the data

```
/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py:2698: DtypeWarning: Columns (25,51) have mixed types. Specify dtype option on import or set low_memory=False.
```

interactivity=interactivity, compiler=compiler, result=result)

I've fixed it by using

```
nfl_data = pd.read_csv("../input/nflplaybyplay2009to2016/NFL Play by Play 2009-2017 (v4).csv", low_memory=False)
sf_permits = pd.read_csv("../input/building-permit-applications-data/Building_Permits.csv", low_memory=False)
```

Options

Tamanna Patil • 10 months ago 2

Great Learning, Rachael. waiting for the coming sessions!!

Options

LakshmipathiKakarla • 10 months ago 2

Really nice and helpful tutorial! Thanks you Rachael!

Options

Faiz Ahmed Khan • 10 months ago 2

An amazing way to learn glad to share my notebook, looking forward to the challenge and hopefully such competitions won't stop coming in https://www.kaggle.com/faizkhan/data-cleaning-challenge-handling-missing-values

Options

Gwin • 10 months ago 2

Excellent job, Rachael ! ... looking forward to the tomorrow's lesson :)

Options

Arunkumar V Ramanan • 10 months ago 1

Wonderful job, Rachael! Well, Done on Day 1 of the 5-Day Data Challenge! Looking forward to getting started Day 2 :)

Options

ANURAGGUPTA • 10 months ago 0

Do we have this exercise in R ?

Options

Rachael Tatman • 10 months ago 1

Hometeamtimeout_Teamposts_Reply

Kannan Veera • 10 months ago 0

Very good learning. Well explained. Thanks!!!. https://www.kaggle.com/kannanv/data-cleaning-challenge-handling-missing-values?scriptVersionId=2929955

Options

John Xu • 10 months ago 0

shafiquekhan • 10 months ago 0

how we can go to R notebook

Options

Replacing NaN with 0 and data cleaning challenge

Xenophontas Psarris • 10 months ago 1

Sunny Chidural • 10 months ago 0

Thanks, Rachel, a very useful lesson
here is my solution for Day1: https://www.kaggle.com/chiduralasunny/data-cleaning-challenge-handling-missing-values

Options

Kier • 10 months ago 0

Gaurav Mukherjee • 10 months ago 1

I'm kind of new to Kaggle. Is there a way to convert the Python notebook to R? I can make all the code changes, but would like all of the text.

Options

Rachael Tatman • 10 months ago 0

shubhamranjan • 10 months ago 0

how many total missing values do we have?

totalcells = np.product(nfldata.shape)

dear kagglers what does that line mean ??

Options

TFu • 10 months ago 3

Wei Wei • 10 months ago 0

I couldn't figure out how to fill the 'Zipcode' column. I think those samples with the same street names and blocks should share the same zipcode. But I failed to implement it. Could you plz help me :(thx

Options

GranVisir • 10 months ago 0

Options

Icaro Bombonato • 10 months ago 0

Mandar Malvé • 10 months ago 0

Nice Job Rachael..looking forward for tomorrow. Here is my notebook
https://www.kaggle.com/mandar2589/data-cleaning-challenge-handling-missing-values/edit/nb?unified=1

Options

Apoorv Patne • 10 months ago 0

I've got a doubt in the last section. Here's my solution :
https://www.kaggle.com/apoorvwatsky/data-cleaning-challenge-handling-missing-v-f07f3f
Can someone please tell me why the missing values in the sf_permits are not replaced by values directly after them?

Options

Betty Holyk • 10 months ago 0

Would you guys give me more comment on this, pl?:

replace all NA's the value that comes directly after it in the same column,

then replace all the reamining na's with 0

subset*nfl*data.fillna(method = 'bfill', axis=0).fillna("0")

Options

Rachael Tatman • 10 months ago 0

Vaughn Shidler • 10 months ago 0

Can someone please explain how the slicing works in the example under "Filling in missing values automatically"? I'm assuming that ["EPA" : "Season"] prints all the columns between "EPA" and "Season," but what does the first colon do? And what if you wanted to print specific rows instead?

Options

Jonas Conde • 10 months ago 3

aseemmyadav • 10 months ago 0

can any one help me where i can write and run the code written above

Options

Chirag Sehra • 10 months ago 1

Sabu Joseph • 10 months ago 0

How can I retrieve the code I saved ? I executed it, versioned, made the code public too

Options

Rachael Tatman • 10 months ago 0

GremlinB • 10 months ago 0

Is there a shutdown procedure, similar to working locally, when we are finished with a day's

work?

Options

Rachael Tatman • 10 months ago 0

Aman Saxena • 10 months ago 0

This was good and easy! Why is it not in R too @Rachael ?
Here's my notebook .. https://www.kaggle.com/amansaxena/data-cleaning-challenge-handling-missing-v-c12d60

Options

Rachael Tatman • 10 months ago 1

SeanGalloway • 10 months ago 0

How do I download the data listed here?

Options

Rachael Tatman • 10 months ago 0

Rishabhjain • 10 months ago 0

Hi Rachel,
I have also signed up for this course but i am not getting any mails regarding this
My mail id is rishabh15virgo@gmail.com
Please have a look

Options

Rachael Tatman • 10 months ago 0

russell luttrell • 10 months ago 0

Thanks for the tutorial. The last cell seems to have its comment cut off mid sentence though.

Options

Rachael Tatman • 10 months ago 0

mgiy • 10 months ago 0

Thank you Rachael!
One question: Is there any special reason for filling the remaining NAs' with string "0", not number 0 in the last line: subset*nfl*data.fillna(method = 'bfill', axis=0).fillna("0") ?

Options

Rachael Tatman • 10 months ago 0

Kaustubh • 10 months ago 0

Thank you for simple and intelligible challenge!
Can you create a similar notebook where the missing values can be filled using various Regression Analysis techniques? In some specific cases, that would be more useful than filling the values with constant or based on a value in the same column of next row.

Options

Rachael Tatman • 10 months ago 0

hklee • 10 months ago 0

Thank you for your clear and really helpful instruction.
How about

subset*nfl*data.fillna(method = 'bfill', axis=0).fillna(method = 'ffill', axis=0)
instead of

subset*nfl*data.fillna(method = 'bfill', axis=0).fillna("0") ?

Options

Rachael Tatman • 10 months ago 0

Siddharth • 10 months ago 0

Hi rachael...Amazing tutorial...really got to learn a lot... I knew the commands but never got to practice on the data sets. Thanks for taking up this topic.
Btw I haven't got the link for Data cleaning challenge, Day 2 yet.
When can I expect to get it? PLEASE LET ME KNOW!

Options

Rachael Tatman • 10 months ago 0

Sami • 10 months ago 0

Hi Rachael, thank you for this tutorial.
Have you ever had issues with columns types ? I remember having issues with columns types because when you convert object type to float, it doesn't handle correctly `NaN` and `None`.

Options

Rachael Tatman • 10 months ago 0

Preety Missirli • 10 months ago 0

Thanks Rachel !
Can you please explain the below :
print("Columns in original dataset: %d \n" % sf*permits.shape[1]) print("Columns with na's dropped: %d" % columns*with*na*dropped.shape[1])

Options

GSD • 10 months ago 0

Rachael Tatman • 10 months ago 0

Rachael Tatman can answer this when she gets back.

Chris • 10 months ago 0

Thanks @Rachael ! I have a question by the way, what's the difference between:

```
subset_nfl_data.fillna(method = 'bfill', axis=0).fillna("0")
subset_nfl_data.fillna(method = 'bfill').fillna("0")
```

Because both give me the same result. Thanks!

Options

LennartGrosser • 10 months ago 1

Chathura • 10 months ago 0

Hi Rachael, I didn't get an email for the second day Challenge?

Options

Rachael Tatman • 10 months ago 0

dennywk2 • 10 months ago

Diana Petuhova • 10 months ago 0

Hi, Rachael!
Why did you use subset*nfl*data.fillna(0), filling missing values with zeros? Is it because the data is already normalised?

Options

Rachael Tatman • 10 months ago 0

SNEHITHA TIGER • 10 months ago 0

Hi@ Rachael Tatman ,
Nice tutorial, Thanks for helping how to code for beginner's like me.
I'm new to python. When I'm trying to run this data an error is shown as initializing from file failed. How to sort it.

Options

Rachael Tatman • 10 months ago 0

Vivi • 10 months ago 0

Hey Rachael, thanks for the tutorial! :D I have one question regarding missing data during training. Suppose there is one real-valued feature that might not exist for every sample (=value missing because it doesn't exist). How do I then train a model if I really want to include this feature(a classical regression model cannot deal with "NaN"s, while trees could perhaps still deal with it)? What methods do you usually choose?

Options

Rachael Tatman • 10 months ago 1

boosting_75 • 10 months ago

Easy and fast way to learn Python;-) Here my code for review:
https://www.kaggle.com/zeus75/data-cleaning-challenge-handling-missing-values

Options

Tyler James Martin • 10 months ago

Another common strategy is to replace with mean.
You can use:
from sklearn.preprocessing import Imputer
imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis = 0)

And then fit & transform your data.

Options

Rachael Tatman • 10 months ago

Abdoulaye Diallo • 10 months ago

Hello, I'm so late for this chalenge. thnk you Rachel. But could someone explain me what .loc[] refer to? and Axis=0?
I'm beginner in python and im still learning basics.

Thanks

Options

Rachael Tatman • 10 months ago

jaishiva • 10 months ago

Hey Rachel is there any way to find out the Street Number Suffix ?

Options

Jan Jeffryes • 10 months ago

David Nichols • 10 months ago

For imputing missing values, NA's can be imputed by first determining the likely distribution of the variable to be imputed, then sampling from its distribution and randomly imputing each NA.
https://epi2020datascience.blogspot.com/2018/01/

Options

Thi An • 10 months ago

Hi all, I'm new to this and I'd like to know if I can Fork this notebook and make it an R network. Thank you very much in advanced.

Options

SNEHITHA TIGER • 10 months ago

```
f
o
r
i
c
(
c
a
t
_
a
d
d
r
e
s
s
[
'
N
U
M
B
E
R
'
]
=
p
d
.
t
o
_
n
u
m
e
r
i
c
(
c
a
t
_
a
d
d
r
e
s
s
[
'
N
U
M
B
E
R
'
]
,
e
r
r
o
r
=
'
i
g
n
o
r
e
'
)
```

Rachael Tatman • 10 months ago 0

blackspider007 • 7 months ago 0

Thanks Rachael for the tutorial! But when I am running the codes in the console here, first the nf1*data did not get defined. I started working with sf*permit but this is even not running. I'm getting an error message "sf*permit is not defined". This is the same error I was getting for nf1*data. Will you please help me?

Options

Rachael Tatman • 7 months ago 0

John Olayisade • 7 months ago 0

Thank you Rachael. Will this work for a ,'R' environment?

Options

Rachael Tatman • 7 months ago 0

Srishti Jaiswal • 7 months ago 0

Thanks a lot for this simple and helpful tutorial. I have a question that when I ran my last code:-

```
"sf_permits.fillna(method='bfill',axis=0).fillna(0)"
```

In the output I got NaN values which was not supposed to come. Can you please help clearing my doubt?

Options

Imran Iskandar • 7 months ago 0

Prabesh123 • 6 months ago 0

Hi Rachel, i am trying my hands on R and was wondering if you run a similar daily challange with R as well ? I could not find it anywhere. Many thanks :)

Options

Rachael Tatman • 6 months ago 0

Bhumika Lamba • 6 months ago 0

Hi,
for the fillna(method='bfill',axis=0).fillna(0) . Why do we need to set axis=0? I am not able to understand how this method works. If it befill method then there won't be any NaN value left to be filled as a zero.

Options

Rachael Tatman • 6 months ago 1

aljqa2 • 6 months ago 0

Hi Rachel -

Thanks - this lesson is awesome! I do have a question - I don't understand why the following head() command doesn't return any rows:

```
rows_with_na_dropped = sf_permits.dropna(axis=0,how = 'any')

rows_with_na_dropped.head()
```

0

My notebook is located @ https://www.kaggle.com/aljqa2/data-cleaning-challenge-handling-missing-values/edit

Thank you!

Options

Joshua • 6 months ago 1

Abdoulaye Diallo • 9 months ago 0

Yo Ram • 6 months ago 0

There are a lot of typo's in the text, but really interesting challenge!

Options

Aniket Gulhane • 5 months ago 0

Hi,

I am new to data science, undergoing a PG diploma.

I have a doubt while treating the NA values. The data I have consist of 15 variable, of which for 2 variables, there are around 988 NA observations out of the 1232 total observations. I am not sure how i should deal with this situation.

Any help is much appreciated!!

Thanks!

Options

Alex Poulain • 5 months ago 0

Thank you! Very nice cleaning tutorial.

Options

Aditya Agrawal • 5 months ago 0

Hi Rachael and everyone here! Thanks a lot for this lesson. I took away some vital points from this notebook- one of the major ones being that every column would need to be dealt with individually! I have just published my first kernel with some pre-processing and EDA on a

clothing-fitting feedback dataset. I would really appreciate it, if you guys could check it out and give me some feedback on it!

Thanks!

Options

Emad Tolba • 4 months ago 0

thanks

Options

Oumaima Hourrane • 4 months ago 0

Very interesting and helpful, thank you!

Options

Ewelina Szymakowicz-Grabania • 4 months ago 0

Thank you Rachel, it is very useful challenge. I am learning a lot.

Options

Farshad • 4 months ago 0

The course is highly appreciated. Because data cleaning is always one of my headaches!

Options

Rachael Tatman • 10 months ago 0

Micheal_Mike • 3 months ago 0

Very helpful

Options

selin • 3 months ago 0

Very helpful Rachael thanks for sharing!

Options

Rachael Tatman • 10 months ago 0

ganesh • 16 days ago 0

how do you handle missing value in categorical data set ?

Options

jors • 8 days ago 0

Thanks for this Rachael!

Options

10 months ago

This Comment was deleted.

Options

10 months ago

This Comment was deleted.

10 months ago

This Comment was deleted.

10 months ago

Options

This Comment was deleted.

10 months ago

Options

This Comment was deleted.

10 months ago

This Comment was deleted.

Options

10 months ago

This Comment was deleted.

10 months ago

This Comment was deleted.

7 months ago

This Comment was deleted.

7 months ago

This Comment was deleted.

Options

7 months ago

This Comment was deleted.

7 months ago

This Comment was deleted.

7 months ago

This Comment was deleted.

Similar Kernels

SQL Scavenger Hunt Handbook

Data Cleaning Challenge: Scale And Normalize Data

SQL Scavenger Hunt: Day 1

Data Cleaning Challenge: Parsing Dates

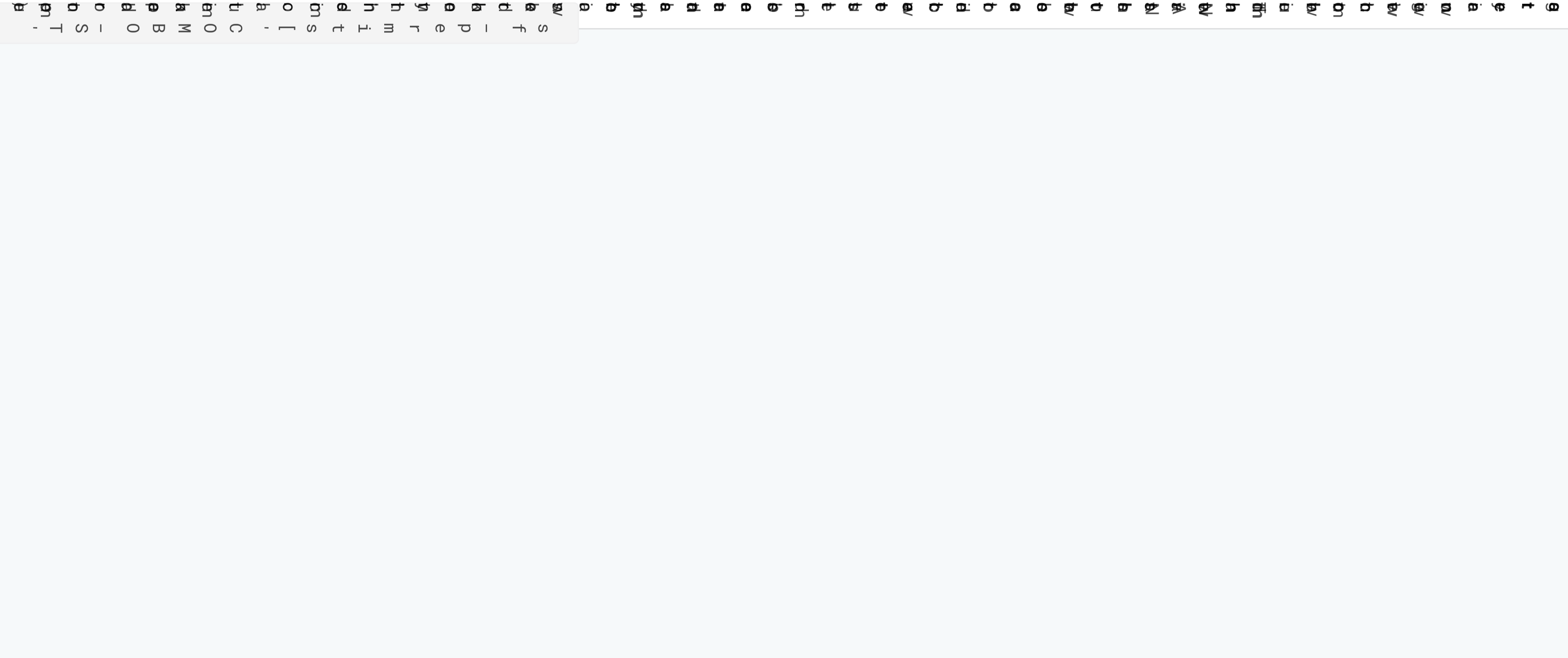
Data Cleaning Challenge: Inconsistent Data Entry

Notebook Data Log Comments

Loading [Contrib]/a11y/accessibility-menu.js

Team Terms Privacy Contact/Support

10 months ago 1



Rachael Tatman (@rachaelxo) · 6 months ago

0

Rachael Tatman (@rachaelxo) · 10 months ago

0

alpha2 · 6 months ago

0

Options

```
(y = lambda x: x**2)
print(y(5))
```

jaiphiva • 10 months ago

2

an apply. eat tubhse (eaerrooaoiyn-gvwaHiusehy)ah. eindgxhmrienet(o) sraeepurrbeanfrdasyillaigdtxlotbbtthphehpha


nengkuan2 • 10 months ago

a h g p m e t s y n o f x e a d w o u l d a g a i n p r o p o s e t o M r. B e n n e t, b u t t h e a d d r e s s o f h i s l e t t e r s e e m e d t o p r e c l u d e a l l h o p e o f p l e a s u r e. I n t h e c o u r s e o f m y l i f e, s o m e p e o p l e h a v e t o l d m e t h a t I h a v e a g r e e a b l e e y e s; a n d t h i s w a s t h o u g h t a n o t h e r c a s e, a n d I w a s t o l d o f t h e p a i n a n d m i s c h i e f t h a t w a s t o b e a v o i d e d b y a c o m p l i a n c e, a n d t h e r e f o r e I w a s e a g e r t o h e a r t h e n a m e o f t h e g e n t l e m a n w h o m I m i g h t b e e x p e c t e d t o l o v e, t h a t I m i g h t b e p r e p a r e d t o s e e h i m w i t h a p r o p e r d e g r e e o f c o o l n e s s. M r. B e n n e t d e l a y e d i t, h o w e v e r, t h a t I d a y m i g h t b e a d d e d t o m y a n x i e t y; s h e w o u l d n o t s a y h e r f a v o u r i t e s u s p i c i o n t h a t I h a d n o r e a s o n t o e x p e c t a n e v e n t f u l a n s w e r.

```
sobre
vivo
esfera
mundo
esfera
(sobre
vivo):
fila
fila
sobre
esfera
[Zip
codes]
(sun
```

s ,] t N h o e % r œ e ï m a $ s ' . o r n o (w b n h a e y b w w [e y , h à N a l u v $ M e œ B S Æ u R f t ' f e] i m x œ F n œ F u l m à f b ç i e ê n r ô d s o – n f l n o d ê a r e r s r e o . s m T t e h (r i n e $ é c i w o s [r m d œ s t w a f [n i ' d R S f a T o l R E r . E s V Y o e , m s] e t r i = e l =

c l
o n
r e
d w
s d
t t
h g
e o
d M
a B
t a
a n
i N
s o
m j
i e
s b
s e
i f
n q
g m
. e
A t
c a
t i
u n
a i
l r
l g
y w
a R
m i
t N
h u
i M
n t
k E
2 R
n l
g t
i m
f i
t d
h h
e t
r a
e l
a u
m e
i s
g o
h d
t j
b i
e a
a s
p t
a e
o :

ttern based on which this data has been provided. But [illegible] so many ways that I'm a

uses some average values (if numerical) or most popular value (if text

```
return a row['Zipcode']
```

Options

s)
h f
a o
m r
e m
d i
t s
o s
p i
o n
s g
t v
i a
t l
h u
e e
r c
e e
. l
l
Options
.
l
Jonas Conde • 10 months ago
t
H h
i e
G r
a e
b a
r h
i a
e n
l d
, y
w
n a
i y
c t
e o
j d
o b
t t
h r
a y
i ?
n
g
Options
t
o
Rachael Tatman • 10 months ag
i
y
u
s
h
g

0

1

ijsdaahbetkoiitedhheellaggooogolliemmakpesaAptthhabbowetevoemmpffttggaanneehabbookkpffetaaddr

e
w
h
e
a
h
t
o
t
e
d
g
e
t
s
q
m
e
a
d
i
o
a
s
s
e
$
(
A
+
l
h
$
t
y
a
g
a
s
h
t
Options
m
o
r
o
w
i
t
h
m
o
r
e
p
a

tieneence.

Options