
DATA ANALYTICS: FORMAT, CLEAN, MANIPULATE DATA 2

Example

- ▶ Consider an address column in this web form data. There is a difference between “I don’t know this person’s address” and “They don’t have an address”.
 - The first sentence can be conveyed using a null
 - The second statement is more appropriately conveyed using an empty text field

P_Id	LastName	FirstName	Address	City
1	Hansen	Ola		Sandnes
2	Svendson	Tove	Borgvn 23	Sandnes
3	Pettersen	Kari		Stavanger



CLEAN Non-printing Characters.

=CHAR(9)&"Monthly report"&CHAR(10)

=CLEAN(A2)



Creating a null value strategy

- ▶ Depending on your dataset and business problem, the ways to deal with null values differ
- ▶ You can:
 - Delete them
 - Fill them with another value - “imputation”
 - Ignore them
 - Find the value



Null value strategies

- ▶ Delete them – dangerous because you may lose many data
 - **Use case:** when you have very few NULL values; you have rows with mostly NULL value

Null value strategies

- ▶ Fill them with another, more meaningful value - “imputation”
 - **Use case:** many rows with a few number of NULL values, you don’t want to throw away all of your data, NULL values skew your aggregation results.
 - **Method:** replace NULL values with the mean or median of the missing value column.




Null value strategies

- ▶ Ignore them
 - **Use case:** when doing quantitative statistics, such as averages and sums; make sure that by doing so, you aren't missing a large percentage of your data.
 - Generally, if over 15% of your data are null, then should raise more questions

Null value strategies

- ▶ Find out what they are; doesn't mean the data doesn't exist somewhere
 - **Use case:** when there is a third-party source we can find the missing data; when there is documentation on how to find null values
 - **Example:** You're doing analysis on bike share data, the weather columns are largely unknown, you could use the date column to turn to 3rd party sources like weather.com to get this data; or launch a campaign to capture this missing data from users
 - **Example:** Some data sets might specifically say "A cell with a -1 element is considered null"



The situation

You've left your job at Kickstarter, and now you're on a contract with the city of New York - they're looking to better understand the 311 requests/complaints so they can re-evaluate staffing needs of PD, FDNY, and paramedics in different locations; but, as we'll see, there is a ton of missing data from this dataset before we'd be comfortable beginning analysis.

Let's look through columns F, G, and H - and how we should approach the NULL values in these particular column(s)



Intro HISTOGRAMS



What is a histogram?

- ▶ A **histogram** shows the distribution of values of a numeric attribute. The x-axis represents the range of values and the y-axis how many data points exist with the given range.
 - get a sense of attribute distribution, range



Difference between a histogram and bar chart

- ▶ bar chart: comparing each student (a *discrete* variable) by score
 - plot *categorical* data
 - [insert a visualization]
- ▶ histogram: no “gaps” between the bars in histograms; used for summarizing, aggregating; describes *what values* a variable takes and also *how often* it takes these values.
 - plot quantitative or *continuous* data
 - [insert a visualization]



How to create bins for a histogram?

- ▶ To create a histogram, you need to divide the data into intervals or **bins**



How to create bins for a histogram?

- ▶ Bins should all be the same size.
- ▶ Bins should include *all* the data, even outliers.
- ▶ Boundaries for bins should land at whole numbers whenever possible.
- ▶ If possible, try to make your data set evenly divisible by the number of bins.
 - Different stories can be told depending on the number of bins you have.