P2: Creating An Analytical Dataset

Understanding Data:

- Data Structure

- Not all the data the we have available to us is easy to use.

- Data can come from different Sources:

Transactional Data: What's recorded at the supermarket for every purchase.

Devices: Data captured from devices we use, like our TV, mobile, cable boxes, etc.

Collected Data: Like weather, census, or flight data.

- Structure of Data:

Data falls into one of 3 categories:

1) Structured
2) Unstructured
3) Semi-Structured

→ Structured Data: Structured data, or data with high degree of organisation. They are typically organized into columns and rows, like in a spreadsheet.

Sometimes columns are called fields and rows are reffred to as records.

Each column represents a variable, and each row represents a record of data.

Structured data are often stored in databases or files, such as spreadsheets, and it's usually easily accessible (accessible). Most importantly, it easy to use.

→ Unstructure Data: They can have no structure to it at all. Since, the data is not organised into a typical columns and rows format, it could be time consuming to work with, as we have to pull what we want out of it.

Eg: A Resume, A Tweet or a Contract document.

→ Semi-Structured Data: is data that has some structure to it, but still requires some work to put it into a structure format of columns and rows.

Eg: A computer log file that requires parsing and manipulating, to put into a format that makes the data easier to analyze.

A catolg of music CD's in an XML file, where we can clearly see a structure, with each having a Title, Artist, Country and other tags, but its not in the typical columns and rows format.

So it needs a bit of work, to get into a format that can be easily used.

DELTA Page No. 24
Date / /

o Introduction to Data Types

• There are two main reasons why data types are important
① When doing calculations across multiple fields.
② When blending data sets.

→ Calculations : Suppose we multiply two no.'s. Multiplication assumes that we're working with numeric fields. So both of the field needs to be number types.

2 x 2
↑ ↑
Numeric
Data type.

→ Data Blending : If we're going to blend two data sets together, we need the field or field's that we're joining on to be the same data type.

Eg:
If we have a customer file and a transaction file, both with the customer ID field, we can join these files together using this field.

Customer file

Customer ID (string)	Name
'1'	Maureen

Transaction file

Customer ID (string)	Transaction
'1'	15

↘ ↙ Join together both files

Customer ID (string)	Name	Transaction
'1'	Maureen	15

We were able to join both the files together because the Customer ID data type was same in both the files. If the Customer ID would have been string in one file and numeric in another, then we cannot join or blend the data together.

- Data Types

Common Field types are :

1) Strings : Strings are any kind of characters, alpha numeric, including symbols.

2) Numeric : Numeric data are no.'s which can be whole no.'s, such as integers, or no.'s with decimal places.

3) Date / Time : can contain a specific date, or a combination of both date and time. This can be really handy for calculating the no. of minutes b/w a caller reporting a problem and its resolution.

4) Boolean Types : also called logical types and is a conditional flag representing either true or false.

5) Special Objects : Objects such as images / maps, reports, objects & sound files, etc.

- Data Issues

- Dirty Data

Dirty data, are data that contain some kind of errors in them, or are in format that are unfriendly or unusable.

Good amount of time need's to be spent cleaning dirty data to make sure to get a correct answer during your analysis.

Eg: If we want to count customers, but some of these customers are represented in duplicate records, then we'll need to fix this before counting them up, or we'll get an inaccurate count.

- Examples of dirty data

  - Data that is not parsed properly, so it appears all in one field instead of multiple fields.

Name		Last Name	First Name
Smith, John	⇒	Smith	John

  - Extra characters in data fields that make it difficult to use the data readily, because those extra characters needs to be removed before analytic can be run.

Name		Name
"John Smith"	⇒	John Smith

- Misspelled data or typos due to human error.

  125 Main Strreet ⇒ 125 Main Street

- Duplicate Data Records.

- Incorrect data, like dates that are from, say, January 1, 1900. This can happen when a user is working with a system that require a value but none is available at the time.

  So, the business practice may be to enter a certain static date that sticks out so it's noticeably wrong, like January 1, 1990.

- Data that does not fit with the expected pattern. For instance, we might have a field containing email addresses, but not all email addresses are in the proper format.

Email
abc@gmail.com
abc@hotmail
xyz@inquisitivecue.in

- Dirty Data - Parsing
Parsing is done when dividing data into parts based on some kind of delimiter.
We can use any delimiter like , (comma), ~ (tilda), | , or any other characters that is not usually found in the data.

- Dealing with Missing Data
There are several options an analyst can use to deal with missing data. The first option is to delete records with missing data.

→ Deleting Missing Data
Deleting missing data is often the default method bcos of its simplicity. No decison needs to be made that might confuse the data. You just get rid of records where there are missing values.
However, we should make sure that deleting data should not have adverse effects on our analysis. For example, if a particular demographic tended to leave a response blank in a survey, then removing records with blank entries will mean that part of population is underpresented.

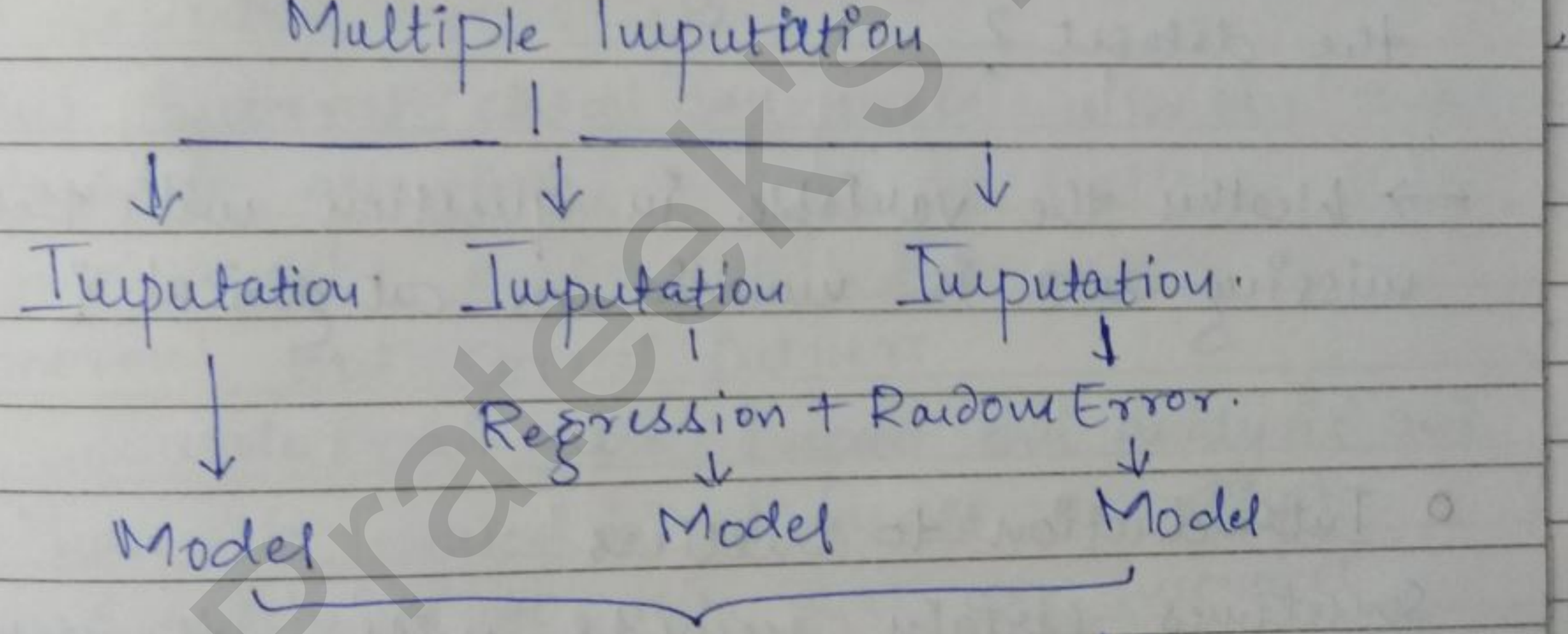
DELTA Page No. 29
Date / /

- Imputation

Another method to deal with missing data is to impute new values in the data that are missing.

Imputation is the process of substituting values in the data where the values are missing, which basically meaning making value up on your own.

- Advanced method of dealing with missing data imputation and deletion are easy to do.
- Multiple Imputation: (Read)



→ We do imputation process multiple times, using the method that will predict the missing value, typically by using regression of other variables and then adding in some random error.
→ With these random sets of data we create the analytical model multiple times and then basically blend those models together.

- Full Information Maximum Likelihood

In this method the missing values aren't actually replaced, but they're handled within the modelling process itself.

This only works for certain types of models and will require special software that can handle this methodology.

- Missing Data factors to Consider

→ How much data is actually missing?

→ How is the missing data distributed across the dataset?

→ Whether the variable in question with the missing data is numeric or categorical?

- Introduction to Outliers

Sometimes certain records within the data just don't seem to fit with the data set as a whole. If it's a numeric field there may be numbers that seem extreme, or way to large or small, compared to the other values in the field.

If it's categorical data it might have values that appears once out of a large set of data.

In statistics, common term for this kind of data is an outlier.

Sometimes, there could be good reasons why this data looks weird and sometimes it maybe bcos the data is just wrong.

DELTA Page No. 31
Date / /

o What is an Outlier ?

The point is not to simply find outlier, but to understand what the extreme data points exist in your data that may affect your analysis.

o Why do we care about Outliers?

Outliers can exist in the data due to two reasons :

i) Incorrect Data
ii) Abnormal but Correct Data

→ Incorrect Data :

Data can be Incorrect due to various reasons, from bad data collection, typos, etc.

This incorrect data can mess with our answers to our business questions.

→ Abnormal but Correct Outliers

Despite accurate can also impact our analysis and we may have to make adjustments accordingly.

So by identifying the data that are incorrect we can correct them, or understand the exceptions that are real.

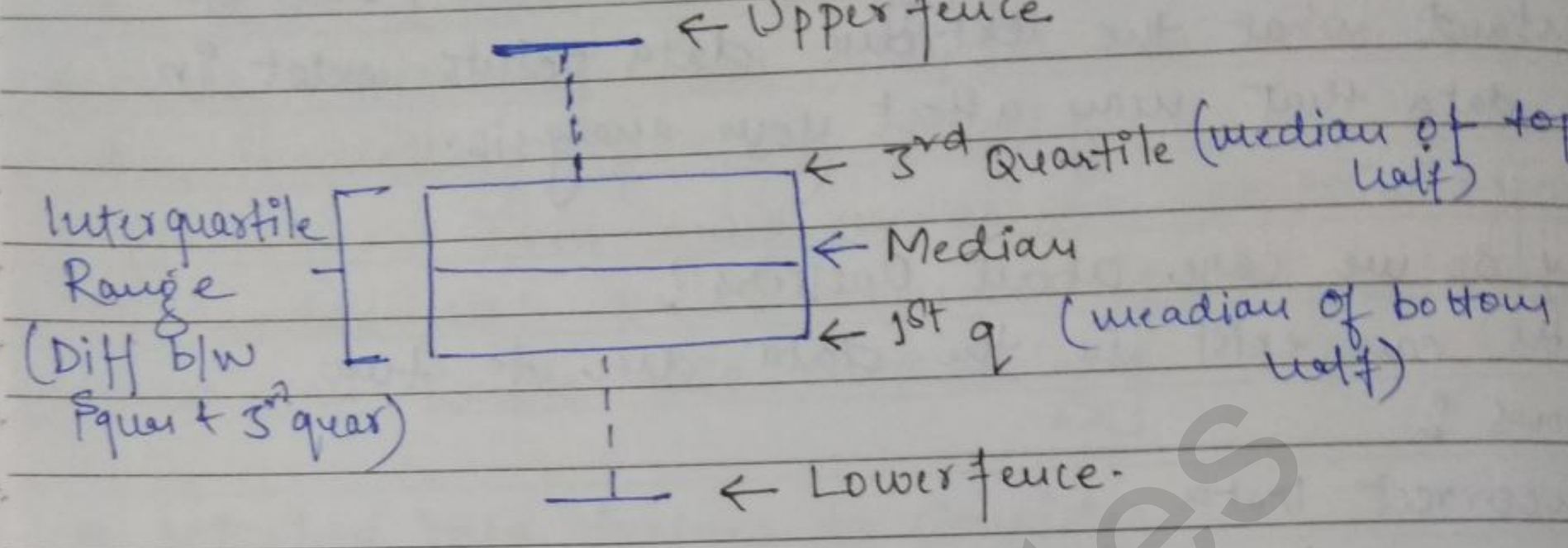
o Identifying Outliers

There are different methods or ways to identify as outliers :

① Box and Whisker plot :

A value needs to be 1.5 times the interquartile range beyond the first and third quartile, then it can be considered outlier.

→ Components of Box and Whiskers plot



→ Steps:

To calculate the upper fence and the lower fence, here are exact steps:

- 1. Calculate 1st quartile Q1 and 3rd quartile Q3 of the dataset. (In Excel we can use function QUARTILE.INC or QUARTILE.EXC)

- 2. Calculate the Interquartile Range:

  IQR = Q3 - Q1.

- 3. Add 1.5 * IQR to Q3 to get the upper fence:

  Upper Fence = Q3 + 1.5 * IQR

- 4. Subtract 1.5 * IQR to Q1 to get the lower fence:

  Lower Fence = Q1 - 1.5 * IQR.

(2) Violin Plot.

DELTA Page No. 33
Date / /

- Data Formatting

- Introduction

Once our data is clean, it might not be formatted properly for us to use in building models.

→ Why format data?

"We might want to build a model forecasting product sales by month over the next 6 months.

If the data we have available is daily product sales, then we're going to need to sum this daily data up for each month".

- Transposing data (Transpose block in Transform tool for Alteryx)

Eg:

Products	1/1/2014	1/2/2014
Skateboard	345	340
Snowboard	2000	1500
Skis	2400	1800

Sales for each month for each product.

But, if we need all the details in one field, we will need to pivot the data orientation for all the months so that it'll display lengthwise.

Products	Name	Sales.
Skateboard	1/1/2014	345
Snowboard	1/1/2014	2000
Skis	1/1/2014	2400
Skateboard	1/2/2014	340
Snowboard	"	1500
Skis	"	800

- Aggregate Data (Summarize Data)

If we have daily sales data for products, but we want to build a forecasting model to forecast monthly sales, then we will need to aggregate the data, so that each record represents a monthly value.

Data	Sales		Month	Sales
1/1/2015	89.02		1	42432.76
1/2/2015	2257.54	⇒	2	39165.11
1/3/2015	697.87		3	43110.24
1/4/2015	2282.91		4	39684.96

- Cross Tabulation:

A Crosstab allows us to take data within a field, and summarize other data to the values within that field to create a matrix.

DELTA Page No. 35
Date / /

Data Blending :

Introduction

- We want to use as much data as is relevant to our analysis, that data may come from diff. places, and as a result, it'll all need to be stitched together into one data file, this is known as data blending.

There are various ways to blend data, two of them are:

1) Unions
2) Joins
3) Fuzzy matching
4) Spatial matching

Union Datasets

- If there are too many records, and exceeds the max. no. of records any one file can handle, then we split the data into multiple files.

- Unioning allows us to take multiple datasets and deal with them as one.

It easier to format or clean up one large dataset than to have to repeat these same operations across multiple datasets, and then bring them together afterwards.

Unioning appends multiple data streams into one unified stream, this makes a longer dataset containing records from both of the original datasets, picture stacking records from one file on top of records from another file, making sure to line up the columns or fields that are common between them.

- Joining Datasets

If we want to use (blend) multiple datasets with different data, we can use joining dataset technique.

For joining different datasets together into one data we will require some field common in every dataset.

- Fuzzy Matching:

Suppose we have two datsets:

Prospects	Customers
↑	↑
List of prospects we bought from a company	List of existing customers

→ We want to reach out to the prospects to turn them into customers, but we don't want to contact people who are already our customers.

So, in this scenario we need to eliminate our existing customers from the prospects lists.

However, given that there probably isn't a nice ID code to match on b/w these two files, then we need to find a way to join the data sets on other fields.

We can use the Name and Address fields, but in practice we don't get a good matches b/w names and addresses representing the same people, because they are frequently spelled slightly differently.

Eg:
Andrew Main, 25 State St ≠ Andy Main, 25 State Street
they both are not gonna join together.
So, this is where Fuzzy Matching comes in.

- Fuzzy matching is really powerful and it'll enable you to join two data sets together where a regular join may fail.

y?

- Fuzzy matching identifies non-identical duplicate data sets by specifying parameters to match on. The values don't have to be exactly the same bcos fuzzy matching uses algorithms to score how similar two words or phrases are.