UNIT - II

Content

- Idea of adding context to the data
- Importance of establishing data context
- Profiling data for context discovery
- How big data affects this effort
- About R
- R and big data

Understanding Your Data Using R

Big Data have the potential to positively impact your marketing efforts, profitability, decision making, or even your life.

There also exists the risk of drawing incorrect conclusions from that same data.

The bigger the data, the bigger the risk

With proper profiling your data, you can see the big picture that your data provides a bit more clearly

Context Clues

A context clue is a **source of information that helps readers understand written content that may be difficult or unique.**

This information offers insight into the content being read or consumed.

Example: “It was an idyllic day: sunny, warm, and perfect…”

Context Clues

With data, context clues should be developed through a process referred to as profiling so that the data consumer can better understand the data when visualized, and also determine what kind of data visualization should be created.

Context or profiling exampleS: calculating the average age of patients or subjects within the data or segmenting the data into time periods (years or months)

Motive for adding Context to data

- To better understand the data when visualized
- To determining what kind of data visualization should be created.
- To gain a new perspective on the data (recognizing and examining a comparison present in the data).
- To make data more relevant (better visualization).

Motive for adding Context to data

Adding context to your data before creating visualizations can certainly make it more relevant for visualization, but context still can't serve as a substitute for value.

There exists many factors such as time of day, or geographic location, or average age etc……

Data visualization needs to benefit those who are going to consume the data.

So, establishing appropriate context - critical requirement.

Data profiling (adding context)

The rule for data profiling: before context, think of a value.

There are **several contextual visualization categories,** which can be used to augment or increase the value and understanding of data for visualization.

These include the following:

- Definitions and explanations
- Comparisons
- Contrasts
- Tendencies
- Dispersion

Definitions and explanations

This is providing additional information or attributes about a data point.

To add to the existing data by creating additional definition or explanatory attributes.

Use existing data points found in the data to create perspectives on the data.

Using patient's weight and height to calculate a new point of data: Body Mass Index (BMI) information

Patient ID	Height	Weight	BMI
10000001	6.2	195	22.60727
10000002	5.9	200	23.76913
10000003	6.0	180	21.2132
10000004	5.1	145	18.51684

Comparisons

This is adding a comparable value to a particular data point.

Example:

1. Total smoking patients visiting a hospital versus the total non-smoking patients visiting a hospital
2. To compare the total number of hospital visits for each state to the average number of hospital visits for a state

State	Cancer Patients	Cancer Patients v National Average
NJ	22	23
PA	21	24
CA	23	29

Contrasts

Adding an opposite to a data point to see if it perhaps determines a different perspective.

An example might be reviewing average body weights for patients who consume alcoholic beverages versus those who do not consume alcoholic beverages:

Avg. Body Weight (Alcohol)	Avg. Body Weight (No Alcohol)
189.0	165.0

Tendencies

These are the typical mathematical calculations (or summaries) on the data as a whole or by other categories within the data, such as mean, median, and mode.

For example, you might add a Median Heart Rate for Age Group that each patient in the data is a member of:

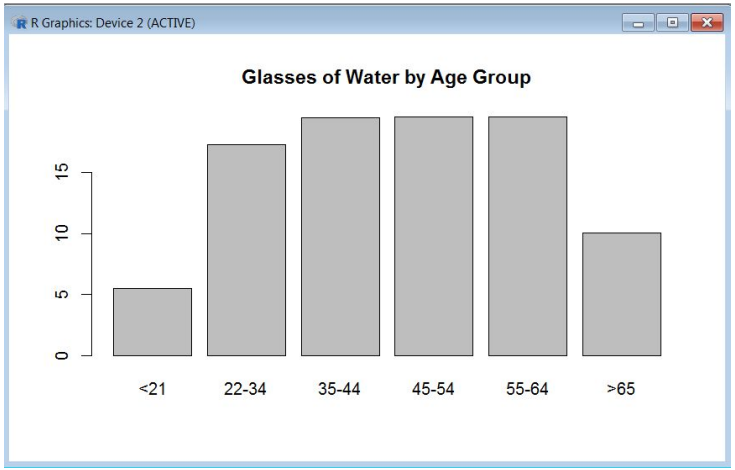
Patient ID	Average Heart Rate	Median Heart Rate for Age Group
10000001	66	71
10000002	100	71
10000003	73	71
10000004	90	71

Tendencies

Example: You might determine what the number of servings of water that was consumed per week by each patient age group

A better approach would be to categorize the data into the age groups.

After we have grouped our data, we can calculate water consumption



Dispersion

These are mathematical calculations (or summaries), such as range, variance, and standard deviation, but they describe the average of a dataset (or group within the data).

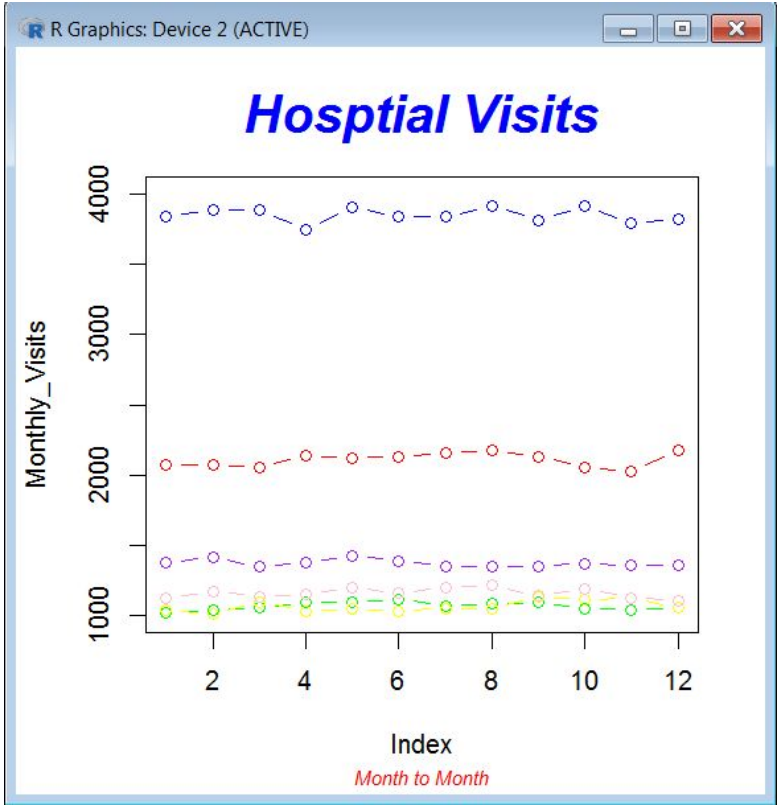
For example, you may want to add the range to a selected value, such as the minimum and a maximum number of hospital stays found in the data for each patient age group:

Patient ID	No Hospital Stays	Hospital Stays Range by age group
10000001	0	0-5
10000002	3	0-5
10000003	2	0-9
10000004	5	0-6

Dispersion

Dispersion measures how various elements selected behave with regards to some sort of central tendency, usually the mean.

For example, we might look ar the total number of hospital visits for each age group, per calendar month in regards to the average number of hospital visits per month



Adding Context (Data profiling)

Is it merely select Insert, then Data Context?

NO, it's not that easy.

Then, how is it done?

The answer is through data profiling.

Data profiling

Data profiling involves logically getting to know about the data through query, experimentation, and review.

Following the profiling process, use the information collected to add context (apply new perspectives) to the data.

Adding context to data requires the manipulation of data to perhaps reformat, adding calculations, aggregations, or additional columns or re-ordering, and so on.

Finally, Data is ready to visualize (or picture).

Profiling Process

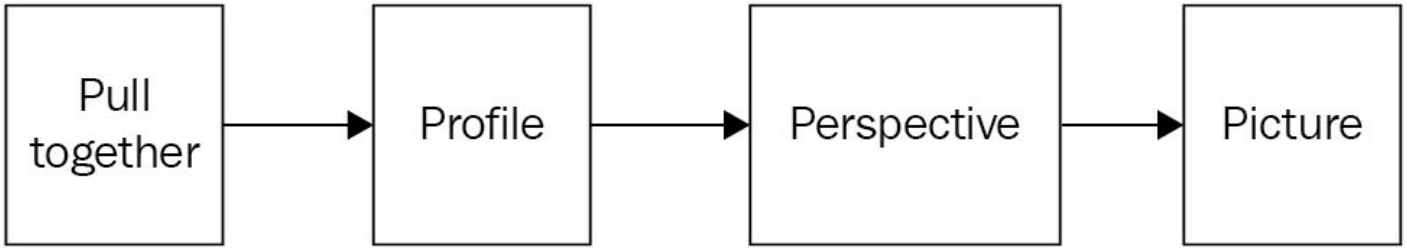
The complete profiling process is as follows:

1. Pull together the data or enough of the data.

2. Profile the data through query, experimentation, and review.

2. Add Perspective(s) or context.

3. Picture (visualize) the data.



About R

R is a language and environment

- easy to learn,
- very flexible in nature, and
- also very focused on statistical computing thus making it great for manipulating, cleaning, summarizing, producing probability statistics, and so on (creating visualizations with your data).

So, R is a great choice for profiling, establishing context, and identifying additional perspectives

About R

Few more reasons to use R when profiling your big data:

- R is used by a large number of academic statisticians, so it's a tool that is not going away.
- R is pretty much platform independent, what you develop will run almost anywhere.
- R has awesome help resources–just Google it; you'll see!

R and Big Data

Although R is free (open sourced), super flexible, and feature rich, but remember that R preserves everything in your machine's memory and this can become problematic when you are working with big data.

R libraries have been developed and introduced that can leverage hard drive space.

Basic Profiling

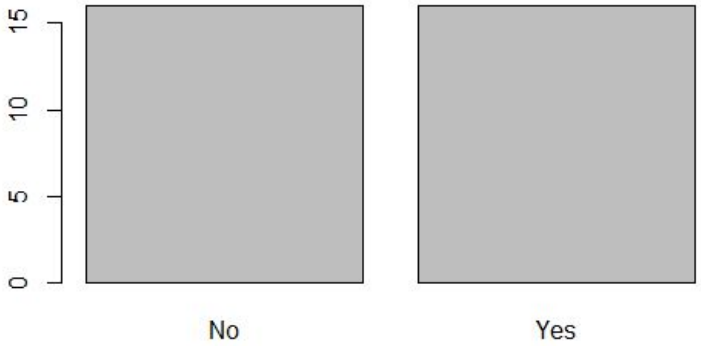
```
> tita.data <- data.frame(Titanic)
> tita.data
   Class     Sex   Age Survived Freq
1    1st    Male Child       No    0
2    2nd    Male Child       No    0
3    3rd    Male Child       No   35
4   Crew    Male Child       No    0
5    1st  Female Child       No    0
6    2nd  Female Child       No    0
7    3rd  Female Child       No   17
8   Crew  Female Child       No    0
9    1st    Male Adult       No  118
10   2nd    Male Adult       No  154
11   3rd    Male Adult       No  387
12  Crew    Male Adult       No  670
13   1st  Female Adult       No    4
14   2nd  Female Adult       No   13
15   3rd  Female Adult       No   89
16  Crew  Female Adult       No    3
17   1st    Male Child      Yes    5
18   2nd    Male Child      Yes   11
19   3rd    Male Child      Yes   13
20  Crew    Male Child      Yes    0
21   1st  Female Child      Yes    1
22   2nd  Female Child      Yes   13
23   3rd  Female Child      Yes   14
24  Crew  Female Child      Yes    0
25   1st    Male Adult      Yes   57
26   2nd    Male Adult      Yes   14
27   3rd    Male Adult      Yes   75
28  Crew    Male Adult      Yes  192
29   1st  Female Adult      Yes  140
30   2nd  Female Adult      Yes   80
31   3rd  Female Adult      Yes   76
32  Crew  Female Adult      Yes   20
```

Head

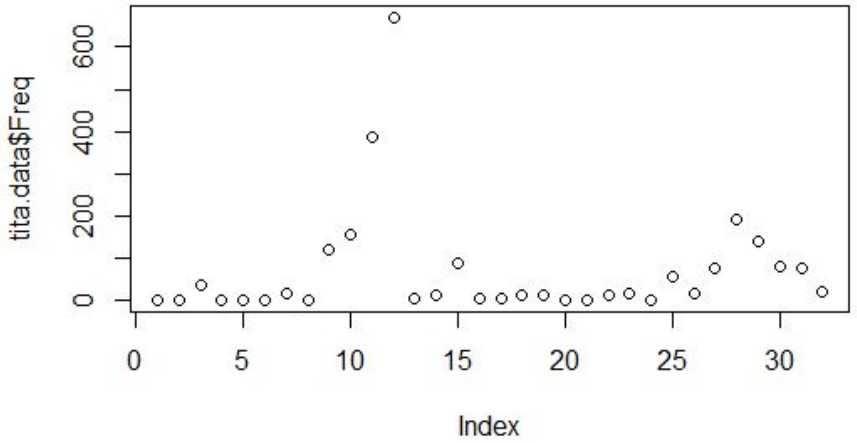
```
> head(tita.data)
  Class    Sex   Age Survived Freq
1   1st   Male Child       No    0
2   2nd   Male Child       No    0
3   3rd   Male Child       No   35
4  Crew   Male Child       No    0
5   1st Female Child       No    0
6   2nd Female Child       No    0
```

Plot

```
> plot(tita.data$Survived)
```

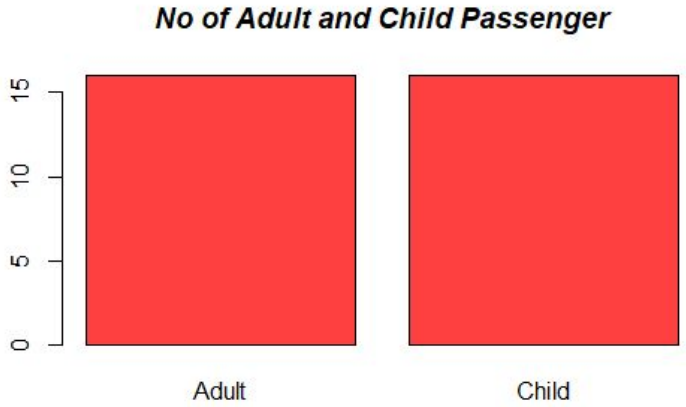



```
> plot(tita.data$Freq)
```



Barplot

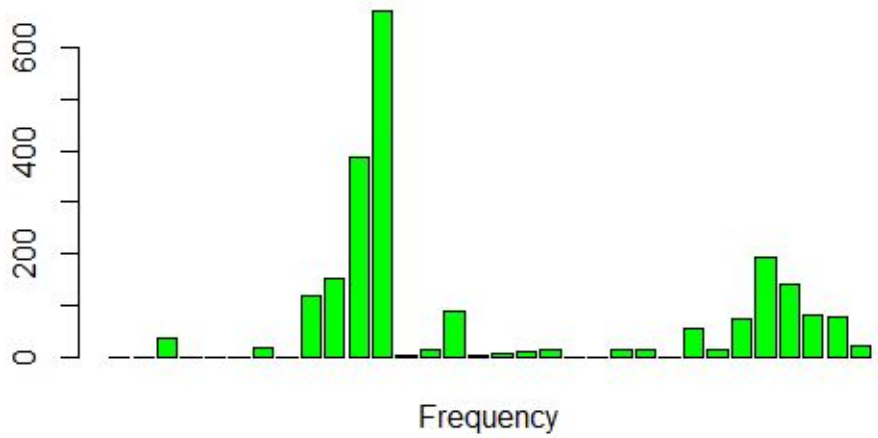
```
> forchart<-ftable(tita.data[,3])
> forchart
 Child Adult
    16    16
```




```
> barnames<-c("Adult","Child")
> barplot(forchart, col = "brown1", border = TRUE,
names.arg = barnames)
> title(main = list("No of Adult and Child Passenger",
font = 4))
```

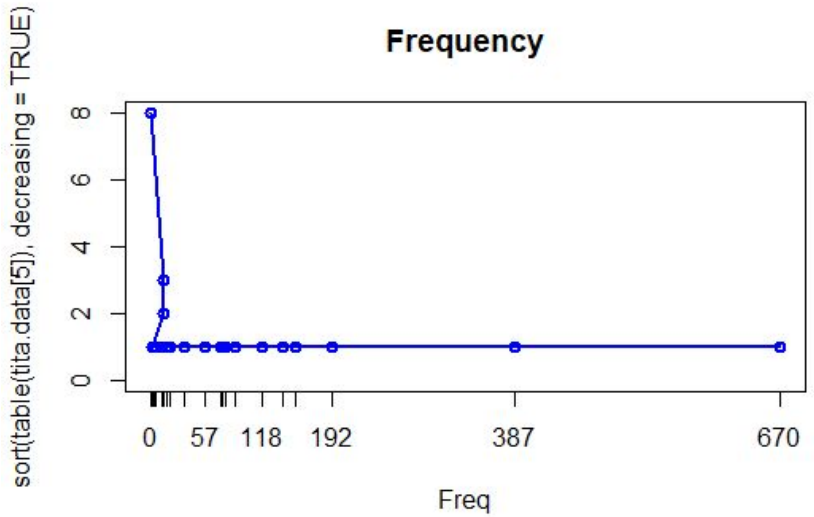
Barplot

```
> barplot(tita.data[,5], col='green', names.arg = "Frequency")
```



Sort

```
> plot(sort(table(tita.data[5]),decreasing = TRUE),type="o", col="blue")
> title(main = list("Frequency", font = 2)
```



Big Data Quality

- Data quality categorized
- DataManager
- DataManager and big data

Programming Language - Data types

Programming languages categorize data into **types or a datatype.**

These categories of data are a defined kind or a set of possible values allowed by the type.

The same concept may be applied to the challenge of data quality.

By understanding the categories of data quality, it makes it easier to identify and address issues with the quality of your big data.

Data Quality Categorized

Garbage In Garbage Out (GIGO).

Computers process all data without judgment.

The quality of data processed by computers is not guaranteed.

If your data is wrong, your results will be wrong.

Data visualizations will only show the value if the data used to create the visualizations has had its quality assured to the appropriate level through routine and regular review and evaluation, practices that, when using large volumes of data, can become extremely demanding.

Data Quality Categorized

Data quality is relative as the level of accurateness or completeness is relative to or relates closely to the intended use of the data.

When considering the level of data quality, one might agree that pollsters routinely determine what level of statistical confidence is required.

They determine the number of people in an entire group and how accurate they want their results to be (accuracy), which then dictates the sampling technique they may use.

Data Quality

"The level of data quality can be affected by the way it is entered, stored, and managed and the process of addressing data quality requires a routine and regular review and evaluation of the data…".

Understanding Categories of Data Quality

Accuracy: There are many **varieties of data inaccuracies** and the most common examples include: poor math, out of range, invalid values, duplication, and more.

Completeness: Data sources may be **missing values** from particular columns, missing entire columns, or even complete transactions.

Update status: You need to establish **the cadence of data refresh**(frequent refresh) or updating as well as have the ability to determine when the data was last saved or updated. This is also **referred to as latency.**

Understanding Categories of Data Quality

Relevance: This involves identification and elimination of information that you don't need or care about, given your objectives.

An example would be removing sales transactions for pickles if you are intending on studying personal grooming products

Consistency: It's common to have to cross-reference or translate information across data sources.

For example, recorded responses to a patient survey may require translation to a single consistent indicator to make later processing or visualizing easier.

Understanding Categories of Data Quality

Reliability: Reliability is chiefly concerned with making sure the method of **data gathering leads to consistent results.**

A common data assurance process involves establishing baselines and ranges and then routinely verifying that data results fall within established expectations.

For example, districts that typically have a mix of both registered Democrat and Republican voters would warrant an investigation if data suddenly was 100% single partied.

Understanding Categories of Data Quality

Appropriateness: Data is considered appropriate if it is **suitable for the intended purpose**; this can be subjective.

For example, it's considered a fact that holiday traffic affects purchasing habits.

Understanding Categories of Data Quality

Accessibility: Data of interest may be watered down in a sea of data you are not interested in, thereby reducing the quality of the interesting data since it is mostly inaccessible.

This is particularly common in big data projects.

Security: Additionally, security may play a role in the quality of your data.

For example, particular computers might be excluded from captured logging files or certain health related information may be hidden and not part of a shared patient data.

DataManager

This is a program that allows you to process and manipulate data in an easy and logical manner through a flexible graphical interface.

DataManager reads from and writes to delimitated files (CSV file), alos supports reading from Open Database Connectivity(ODBC).

Data Manager allows you to construct scenes of conceptual designs using simple mouse clicks.

These scenes describe how your data will be processed and transformed.

All of the scenes you create can be saved and reused.

DataManager

DataManager makes use of the concept of functional nodes.

With these nodes, you form a design by adding various nodes and linking them, such that the links form the flow of your data processing.

Each DataManager node performs a single function on your data and once it completes that function, it passes your data to the node it is linked to.

Can use DataManager to create very straightforward designs or very complicated designs.

DataManager - Node Functionalities

Node functionalities available in DataManager include appending, deriving, distinction, fill, filter, merge, sample, select, and sort.

Output options include distribution, histogram, database (DB), ODBC, quality, statistics, table, and XY plotting.

DataManager and Big data

DataManager can handle very large datasets or files.

When it comes to big data, it has essentially constrained your machine resources–processor speed, memory, and storage space.

With features, functionalities and appropriate strategies of Data Manage, you can overcome some of the big data challenges and limitations a machine may apply

DataManager Software

Data Manager Software download link:

https://datamanager.com.au/

Assignments

Data Profiling (Assignment - 1):

For your dataset, do the different profiling process to look at the data using techniques like defining new data points based upon the existing data, performing comparisons, looking at contrasts (between data points), identifying tendencies, and using dispersions to establish the variability of the data using R.

Addressing Quality using Data Manage (Assignment - 2):

Using Data Manager create a scene and add quality to your dataset addressing consistency,reliability, appropriateness, update status, relevance and completeness.

THE END