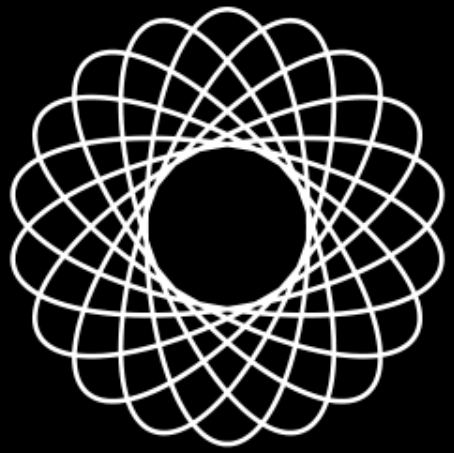
DATA SCIENCE



BUMPER

DATA SCIENCE

EXPLORATORY DATA ANALYSIS

★ Data Exploration Part 1 ★

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

- What is data?

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

- What is data?
- I have a data set, can I start modeling?

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

- What is data?
- I have a data set, can I start modeling?
- Good or Bad : How do I assess quality of data?

EDA - Overview

Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

- What is data?
- I have a data set, can I start modeling?
- Good or Bad : How do I assess quality of data?
- Data Characteristics – How to summarize information

EDA - Overview

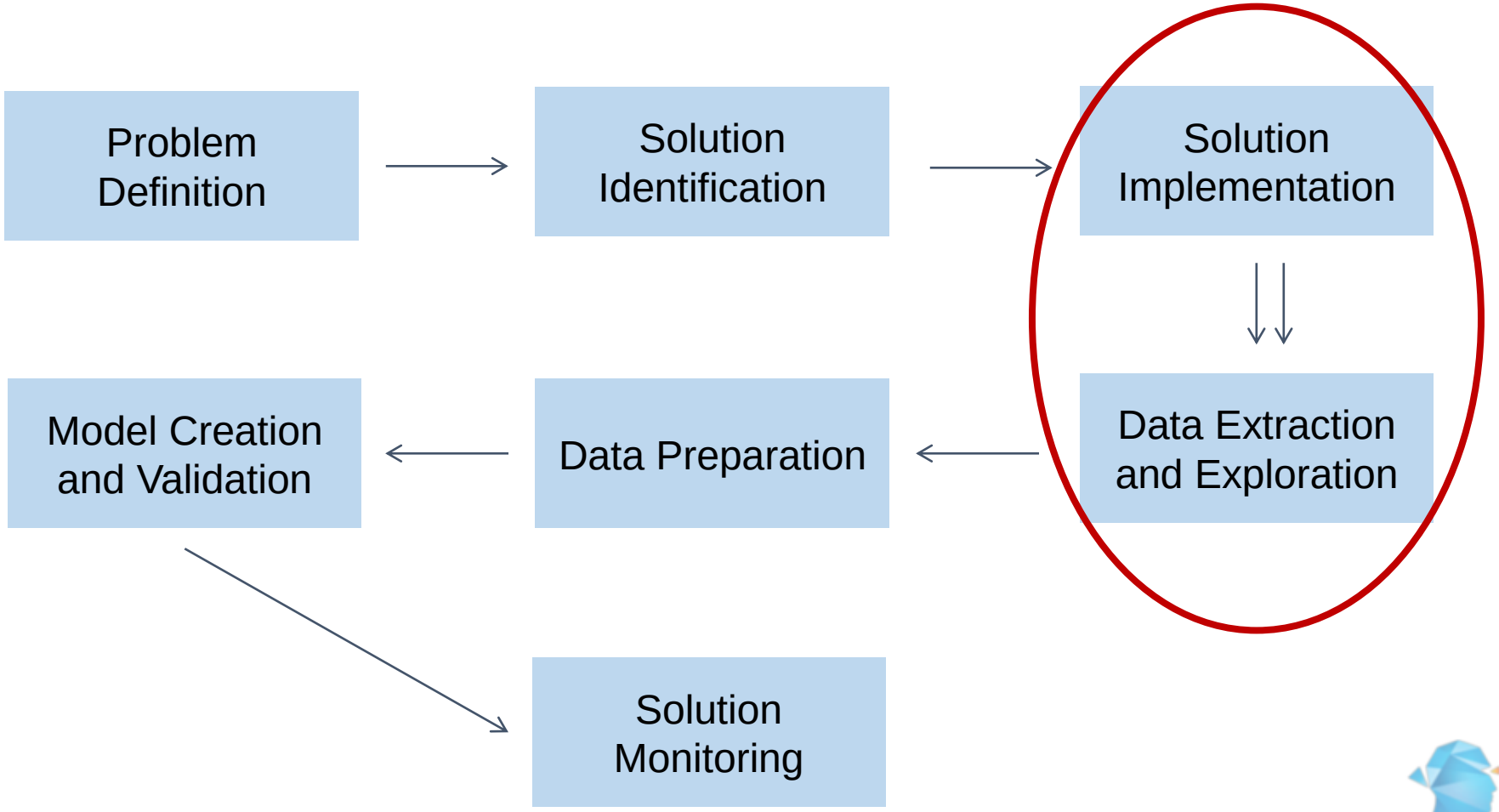
Exploratory Data Analysis is the assessment of the quality and characteristics of data available to tackle business problems

We will cover the following in the session:

- What is data?
- I have a data set, can I start modeling?
- Good or Bad : How do I assess quality of data?
- Data Characteristics – How to summarize information
- Data not conforming to expectations – big problem?

Data Exploration

Where in the analytics methodology?



Data Pre-processing

Data exploration and data preparation usually are done together

Data Pre-processing

Data exploration and data preparation usually are done together

Data Extraction
Data Integration
Data Assessment

Data Exploration

Data Pre-processing

Data exploration and data preparation usually are done together

Data Extraction
Data Integration
Data Assessment

→ Data Exploration

Data Cleaning
Data Transformation
Data Reduction

→ Data Preparation

Data Exploration Case Study

A mobile service provider has noticed a lot of attrition in customers subscribing to their services. It wants to understand what is driving attrition, and identify potential options to retain customers

Data Exploration Case Study

A mobile service provider has noticed a lot of attrition in customers subscribing to their services. It wants to understand what is driving attrition, and identify potential options to retain customers

Data available includes:

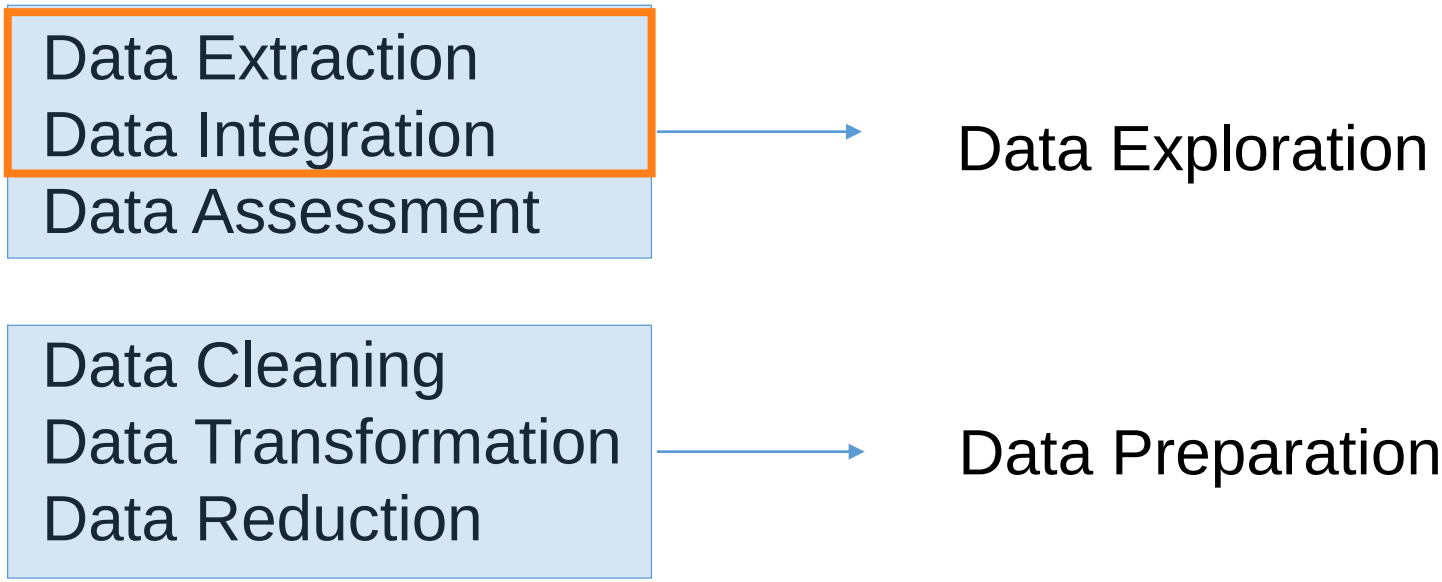
- Subscriber information: including age, location
- Service start date
- Service end date
- Usage by month in minutes
- Plan details
- Promotion details

Data Exploration Case Study

sbscrp_id	minuse1	minuse2	minuse3	minuse4	Plan Type	prom1	prom2	prom3	prom4	svc_start_dt	svc_end_dt	BIRTH_DT	zip_code	NEW_CELL_IND
19164958	57	21	40	60	200 for 10	0	0	0	0	09-Aug-01	09-Apr-03	19730120	30339	Y
39244924	80	510	173	139	200 for 10	0	0	1	1	30-Oct-01	26-Mar-02	19730418	48125	N
39578413	439	805	874	1133	Nights and Weekends	0	0	0	0	16-Aug-01	27-Apr-02	19820820	63120	Y
40992265	200	304	29	135	Nights and Weekends	0	1	0	0	11-Oct-01	26-May-02	19500915	94022	Y
43061957	245	244	286	238	Nights and Weekends	0	0	0	0	09-Aug-01	22-Oct-02	19331224	11212	U
47196850	27	175	91	221	200 for 10	0	0	0	0	19-Oct-01		19730930	21221	N
51236987	77	549	464	256	Nights and Weekends	0	0	0	0	19-Oct-01	06-Feb-03	19760409	80919	N
51326773	131	274	438	320	Nights and Weekends	0	1	0	0	16-Aug-01	11-Apr-02	19480621	95828	U
54271247	37	56	60	72	200 for 10	0	0	0	0	27-Aug-01	30-Nov-02	19570713	75482	U
70765025	169	128	35	0	Nights and Weekends	0	1	1	0	02-Aug-01	26-Nov-02	19761001	8618	Y
70781923	311	334	409	261	Nights and Weekends	0	1	1	1	02-Aug-01		19830305	65201	Y
70782614	2	177	280	177	Nights and Weekends	0	0	0	0	30-Aug-01		19780407	57110	U
70797166	83	217	202	181	Nights and Weekends	0	1	1	1	05-Aug-01		19500411	7018	U
70992941	126	247	428	409	Nights and Weekends	0	1	0	0	02-Aug-01	30-Aug-02	19340901	92543	U
70995813	163	350	213	426	Nights and Weekends	0	1	0	0	02-Aug-01		19821119	35214	U
71000813	251	275	356	371	Nights and Weekends	0	1	1	0	02-Aug-01		19521214	6525	N
71001054	46	116	48	69	200 for 10	0	1	1	1	02-Aug-01	23-Jun-02	19580302	27504	U
71008771	144	74	73	107	Nights and Weekends	0	1	1	1	02-Aug-01	23-Jun-02	19550116	78852	U
71011165	207	105	2	113	Nights and Weekends	0	0	0	0	02-Aug-01	11-Oct-02	19750609	7865	U
71014528	347	334	283	240	Nights and Weekends	0	1	0	0	02-Aug-01		19471106	30215	U
71015314	80	46	49	45	200 for 10	0	0	0	0	02-Aug-01	22-Feb-03	19760415	48509	U

Data Pre-processing

Data exploration and data preparation usually are done together



Data Exploration : Step 1

Sometimes, the first step of data exploration will be data extraction

Data Exploration : Step 1

Sometimes, the first step of data exploration will be data extraction

- Data maybe in database table or flat file format

Data Exploration : Step 1

Sometimes, the first step of data exploration will be data extraction

- Data maybe in database table or flat file format
- May need specialized queries to extract the right data

Data Exploration : Step 1

Sometimes, the first step of data exploration will be data extraction

- Data maybe in database table or flat file format
- May need specialized queries to extract the right data
- If data is already available, then may need to load the data into the specified tool

Data Exploration : Step 2

After data extraction, another intermediate step – data integration

Data Exploration : Step 2

After data extraction, another intermediate step – data integration

- Usually, all the raw data required for analysis may not be part of a single database or table (transactions data and customer data)

Data Exploration : Step 2

After data extraction, another intermediate step – data integration

- Usually, all the raw data required for analysis may not be part of a single database or table (transactions data and customer data)
- Sometimes, we may also need to add data sourced from third-party sources to make the analysis robust

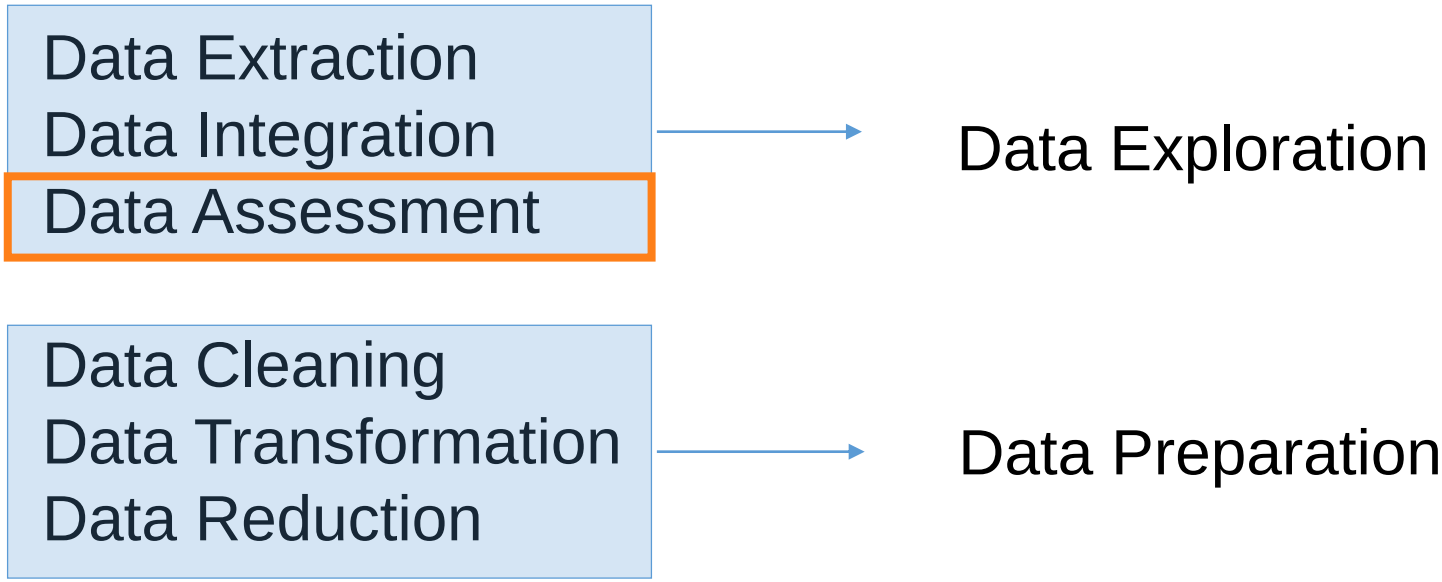
Data Exploration : Step 2

After data extraction, another intermediate step – data integration

- Usually, all the raw data required for analysis may not be part of a single database or table (transactions data and customer data)
- Sometimes, we may also need to add data sourced from third-party sources to make the analysis robust
- The data integration may require specialized skills especially if data is not all at the same level of aggregation

Data Pre-processing

Data exploration and data preparation usually are done together



Data Exploration : Step 3

Data Assessment:

Data Exploration : Step 3

Data Assessment:

1. What is the information contained in the data?

Data Exploration : Step 3

Data Assessment:

1. What is the information contained in the data?

2. What is the quality of this information?

Data Exploration : Step 3

Data Assessment:

1. What is the information contained in the data?
2. What is the quality of this information?
3. Is the data complete?

Data Exploration : Step 3

Data Assessment:

1. What is the information contained in the data?
2. What is the quality of this information?
3. Is the data complete?

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

So, What is “Data?”

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

So, What is "Data?"

Thesaurus definition: "**Information in visible form**"

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

So, What is "Data?"

Thesaurus definition: "**Information in visible form**"

- **Information – any kind**

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

So, What is "Data?"

Thesaurus definition: "**Information in visible form**"

- **Information – any kind**
- **Visibility – collected and compiled, accessible**

EDA – Preliminary Preliminaries

Analytical approaches to business problems focus on data driven analysis and conclusions.

So, What is "Data?"

Thesaurus definition: "**Information in visible form**"

- **Information – any kind**
- **Visibility – collected and compiled, accessible**
- **Neutral, Factual – always**

EDA – Preliminary Preliminaries

How to collect Data? Why is it important to understand mode of data collection?

EDA – Preliminary Preliminaries

How to collect Data? Why is it important to understand mode of data collection?

- **Primary v/s Secondary**

EDA – Preliminary Preliminaries

How to collect Data? Why is it important to understand mode of data collection?

- **Primary v/s Secondary**
- **Manual v/s Automated**

EDA – Preliminary Preliminaries

How to collect Data? Why is it important to understand mode of data collection?

- **Primary v/s Secondary**
- **Manual v/s Automated**
- **Census v/s Sample**

EDA – Data Types

Data Classification

EDA – Data Types

First Step:

Data Classification

EDA – Data Types

First Step:

Data Classification

Qualitative :

Quantitative :

EDA – Data Types

First Step:

Data Classification

Qualitative :

Name, Address, Gender (M/F)

Quantitative :

EDA – Data Types

First Step:

Data Classification

Qualitative :

Name, Address, Gender (M/F)

Quantitative :

Discrete : Nominal / Ordinal

Continuous

Date / Time

EDA – Data Types

Other classification schemes

EDA – Data Types

Other classification schemes

- Primary v/s Secondary

EDA – Data Types

Other classification schemes

- Primary v/s Secondary
- Actual v/s Derived

EDA – Data Types

Other classification schemes

- Primary v/s Secondary
- Actual v/s Derived
- Based on usefulness

EDA Case Study : Data Classification

Data contained is of three types:

Variable	Type
sbscrp_id	Quantitative
minuse1	Quantitative
minuse2	Quantitative
minuse3	Quantitative
Plan_type	Qualitiative
prom1	Quantitative
prom2	Quantitative
prom3	Quantitative
svc_start_dt	Date
svc_end_dt	Date
BIRTH_DT	Quantitative
zip_code	Quantitative
NEW_CELL_IND	Qualitiative

EDA Case Study : Data Classification

Data contained is of three types:

a. **Qualitative** – Type of plan, Type of promotion

Variable	Type
sbscrp_id	Quantitative
minuse1	Quantitative
minuse2	Quantitative
minuse3	Quantitative
Plan_type	Qualitiative
prom1	Quantitative
prom2	Quantitative
prom3	Quantitative
svc_start_dt	Date
svc_end_dt	Date
BIRTH_DT	Quantitative
zip_code	Quantitative
NEW_CELL_IND	Qualitiative

EDA Case Study : Data Classification

Data contained is of three types:

a. **Qualitative** – Type of plan, Type of promotion

b. **Quantitative** – Subscriber Ids, Zip Codes, Number of minutes used in month

Variable	Type
sbscrp_id	Quantitative
minuse1	Quantitative
minuse2	Quantitative
minuse3	Quantitative
Plan_type	Qualitiative
prom1	Quantitative
prom2	Quantitative
prom3	Quantitative
svc_start_dt	Date
svc_end_dt	Date
BIRTH_DT	Quantitative
zip_code	Quantitative
NEW_CELL_IND	Qualitiative

EDA Case Study : Data Classification

Data contained is of three types:

a. **Qualitative** – Type of plan, Type of promotion

b. **Quantitative** – Subscriber Ids, Zip Codes, Number of minutes used in month

c. **Date** – Date denoted separately. It is quantitative data, but needs to be treated with care

Variable	Type
sbscrp_id	Quantitative
minuse1	Quantitative
minuse2	Quantitative
minuse3	Quantitative
Plan_type	Qualitiative
prom1	Quantitative
prom2	Quantitative
prom3	Quantitative
svc_start_dt	Date
svc_end_dt	Date
BIRTH_DT	Quantitative
zip_code	Quantitative
NEW_CELL_IND	Qualitiative

What is Data Exploration?

What is Data Exploration?

1. What is the information contained in the data?

What is Data Exploration?

1. What is the information contained in the data?

2. What is the quality of this information?

What is Data Exploration?

1. What is the information contained in the data?
2. What is the quality of this information?
3. Is the data complete?

Information contained in the data?

1. How was the data collected?

Information contained in the data?

1. How was the data collected?

2. Are the fields accurately labeled?

Information contained in the data?

1. How was the data collected?
2. Are the fields accurately labeled?
3. Is there any missing information?

What is Data Exploration?

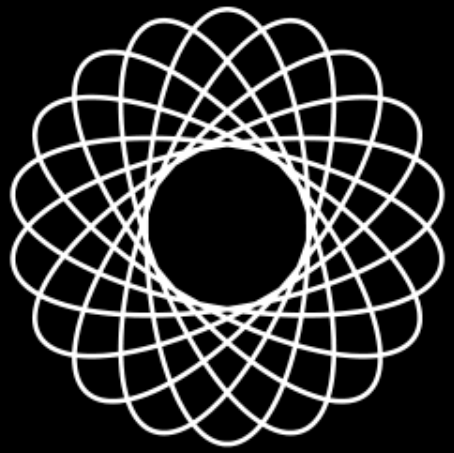
1. What is the information contained in the data?
2. **What is the quality of this information?**
3. Is the data complete?

Recap

Data exploration, the need for it and how to look at information contained in the data

THANK YOU

DATA SCIENCE



BUMPER

DATA SCIENCE

EXPLORATORY DATA ANALYSIS

★ Data Exploration Part 2 ★

What is Data Exploration?

1. What is the information contained in the data?
2. **What is the quality of this information?**
3. Is the data complete?

Data Exploration - Steps

Data integrity, usefulness cannot be assumed

Data Exploration - Steps

<u>Data integrity, usefulness cannot be assumed</u>

Basic sanity checks – do I see what I expect to see?

Data Exploration - Steps

<u>Data integrity, usefulness cannot be assumed</u>

Basic sanity checks – do I see what I expect to see?

- Should I always see what I expect to see?

Data Exploration - Steps

<u>Data integrity, usefulness cannot be assumed</u>

Basic sanity checks – do I see what I expect to see?

- Should I always see what I expect to see?
- Anomalies – always noise? What is an anomaly?

Data Exploration - Steps

Data integrity, usefulness cannot be assumed

Basic sanity checks – do I see what I expect to see?

- Should I always see what I expect to see?
- Anomalies – always noise? What is an anomaly?
- Domain knowledge

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

1. How was it collected?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

1. How was it collected?

2. Is it the universe?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

1. How was it collected?

2. Is it the universe?

3. Active or Passive?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

1. How was it collected?

2. Is it the universe?

3. Active or Passive?

4. What are values in each variable?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

5. Do you understand the values?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

5. Do you understand the values?

6. Is there missing data?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

5. Do you understand the values?

6. Is there missing data?

7. Do you see unexpected values?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

5. Do you understand the values?

6. Is there missing data?

7. Do you see unexpected values?

8. Is the data enough? Do you need more variables?

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

- Once we review the data, we can formulate an approach to answer the problems posed by the telecom service provider

EDA Case Study : Quality of Data

What data is contained in the telecom dataset?

- Once we review the data, we can formulate an approach to answer the problems posed by the telecom service provider

- Without a thorough review, we run the danger of applying techniques that may not be appropriate, leading to incorrect results

Assessing Data Quality

6 steps in data exploration

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative

2. Generate derived variables

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics
4. Cross tabulation

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics
5. Graphical analysis
4. Cross tabulation

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics
4. Cross tabulation
5. Graphical analysis
6. Anomaly detection

EDA Case Study Steps

1. **Transform qualitative data to quantitative**

2. Generate required derived values

3. Generate summary statistics

    - Big Picture view
    - Basic sanity checks

EDA Case Study Steps

1. **Transform qualitative data to quantitative**

Plan type currently has : 200 for 10, Nights and Weekends, and Coast to Coast.

Create a variable Plan_Type1 = 1 if field has “200 for 10”; Plan_Type2 = 1 if field has “Nights and Weekends” etc.

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. **Generate required derived values**

EDA Case Study Steps

Derived Variables

Is there a direct variable for the objective in this dataset?

EDA Case Study Steps

Derived Variables

Is there a direct variable for the objective in this dataset?

Attrition?

If service end date exists, customer has left. Create a variable that lists customer left = 1 if service date > 0

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values

Other examples of derived variables

- Age
- Birthdt not useful directly
- Others?

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values

More complex

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values

More complex

Bucket users into High Medium Low

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values

More complex

Bucket users into High Medium Low

Judgment and domain knowledge is required

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate Summary Statistics

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate Summary Statistics

Generate summary statistics

Basic summary statistics help identify values in data, range, missing value issues, and potential outliers

Data Exploration Summary Statistics

1. **Summary statistics to assess**
    - Completeness of data
    - Missing data
    - Outliers

Data Exploration Summary Statistics

1. **Summary statistics to assess**
    - Completeness of data
    - Missing data
    - Outliers

2. **Why “summary” stats?**
    - Large data sets

Data Exploration Summary Statistics

1. **Summary statistics to assess**
    - Completeness of data
    - Missing data
    - Outliers

2. **Why “summary” stats?**
    - Large data sets

3. **Which summary stats?**
    - Min, Max
    - mean, median, mode
    - Std. Deviation
    - # Missing observations

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate summary statistics

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate summary statistics

    3a. Generate understanding of “big picture”

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate summary statistics

    3a. Generate understanding of “big picture”

    3b. Basic sanity checks

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate summary statistics

    3a. Generate understanding of “big picture”

    3b. Basic sanity checks

    - What % of customers have left?
    - What is the earliest start date? Latest end date?
    - If customer has left, is there usage date post leaving?
    - What is the maximum monthly usage in minutes? Minimum? Are there negative numbers?

EDA Case Study Steps

1. Transform qualitative data to quantitative
2. Generate required derived values
3. Generate summary statistics

    3a. Generate understanding of “big picture”

    3b. Basic sanity checks

    - What % of customers have left?
    - What is the earliest start date? Latest end date?
    - If customer has left, is there usage date post leaving?
    - What is the maximum monthly usage in minutes? Minimum? Are there negative numbers?

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Initial Findings:

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Initial Findings:

1. Unexpected values: negative for minutes used

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Initial Findings:

1. Unexpected values: negative for minutes used
2. Very high monthly minutes?

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Initial Findings:

1. Unexpected values: negative for minutes used
2. Very high monthly minutes?
3. Should all zip codes have standard length

EDA Case Study : Basic Checks, Missing Values, Outliers

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Initial Findings:

1. Unexpected values: negative for minutes used
2. Very high monthly minutes?
3. Should all zip codes have standard length
4. What do the means of the promotion variables reveal?c0

EDA Case Study : Basic Checks

Variable: Minuse 2			
Lowest Value	Record Number	Highest Value	Record Number
-55	851	1500	6211
0	12495	1500	6416
0	12325	1500	7040
0	12320	1500	7525
0	12258	1500	10791

Only one negative number

Potential data entry error?

Only one huge number

Potential data entry error?

Variable: Minuse 4			
Lowest Value	Record Number	Highest Value	Record Number
0	12495	1389	3165
0	12487	1500	2659
0	12476	1500	4458
0	12473	1500	4476
0	12459	177700	154

Many instances of < 5 digit codes

Need further investigation

Variable Zip Code		
Zip Code Length	Number of observations	% of Observations
3 digit zip code	75	0.6
4 digit zip code	922	7.38
5 digit zip code	11503	92.02

EDA Case Study : Qualitative Variables

Plan Type	Frequency	Percent
200 for 10	6643	53.14
Coast to Coast	941	7.53
Nights & Weekends	4916	39.33

NEW_CELL_IND	Frequency	Percent
N	672	5.38
U	9875	79
Y	1953	15.62

EDA Case Study : Qualitative Variables

Plan Type	Frequency	Percent
200 for 10	6643	53.14
Coast to Coast	941	7.53
Nights & Weekends	4916	39.33

NEW_CELL_IND	Frequency	Percent
N	672	5.38
U	9875	79
Y	1953	15.62

1. 200 for 10 most popular
2. Coast to Coast least popular
3. Transform to numeric variable:
   - In excel – find and replace
   - In SAS – create variable

1. Three values for new cell, is that expected?
2. “U” is most frequent, so does “U” denote No?
3. Transform to numeric variable:
   - In excel – find and replace
   - In SAS – create variable

EDA Case Study : Derived Variables

Customer Status	Frequency	Percent
Retained	7733	61.86
Lost	4767	38.14

Deriving actual customers lost reveals almost 40% of total customers have ended service – big number

EDA Case Study : Derived Variables

Customer Status	Frequency	Percent
Retained	7733	61.86
Lost	4767	38.14

Deriving actual customers lost reveals almost 40% of total customers have ended service – big number

Age Group	Frequency	Percent
< 25	1861	14.89
25 - 35	3003	24.02
35 - 45	2539	20.31
45 - 55	2618	20.94
> 55	2479	19.83

EDA Case Study : Derived Variables

Customer Status	Frequency	Percent
Retained	7733	61.86
Lost	4767	38.14

Deriving actual customers lost reveals almost 40% of total customers have ended service – big number

Age Group	Frequency	Percent
< 25	1861	14.89
25 - 35	3003	24.02
35 - 45	2539	20.31
45 - 55	2618	20.94
> 55	2479	19.83

A distribution of customers by age group shows no skew by age – is that reasonable or contrary to expectations?

Assessing Data Quality

6 steps in data exploration

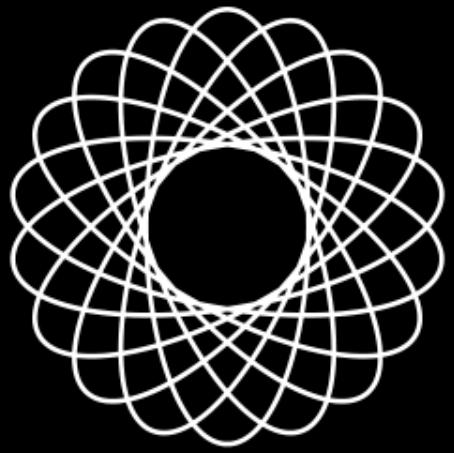
1. Transform qualitative into quantitative ✓
2. Generate derived variables ✓
3. Summary statistics ✓
4. Cross tabulation
5. Graphical analysis
6. Anomaly detection

Recap

Transform qualitative data into quantitative data, generate derived variables and summary statistics

THANK YOU

DATA SCIENCE



BUMPER

DATA SCIENCE

EXPLORATORY DATA ANALYSIS

★ Data Exploration Part 3 ★

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative ✓
2. Generate derived variables ✓
3. Summary statistics ✓
4. Cross tabulation
5. Graphical analysis
6. Anomaly detection

Data Exploration – Cross-Tabs

Investigating patterns one variable at a time is a starting point, but need to look for patterns across groups of variables

Data Exploration – Cross-Tabs

Investigating patterns one variable at a time is a starting point, but need to look for patterns across groups of variables

For example, age buckets created by looking at :

Proportion of attrition by age group

Data Exploration – Cross-Tabs

Investigating patterns one variable at a time is a starting point, but need to look for patterns across groups of variables

For example, age buckets created by looking at :

Proportion of attrition by age group

The simplest way to look at patterns across groups of variables is to create **cross-tabs**

Data Exploration – Cross-Tabs

Age Bucket	Retained	Lost	Retained %	Lost %
< 25	869	992	47%	53%
25 - 35	1651	1352	55%	45%
35 - 45	1555	984	61%	39%
45 - 55	1771	847	68%	32%
> 55	1887	592	76%	24%

Data Exploration – Cross-Tabs

Age Bucket	Retained	Lost	Retained %	Lost %
< 25	869	992	47%	53%
25 - 35	1651	1352	55%	45%
35 - 45	1555	984	61%	39%
45 - 55	1771	847	68%	32%
> 55	1887	592	76%	24%

Clearly, high attrition for customers below the age of 25, and much lower for customers above the age of 55.

- Potential to create age buckets as less than 25, between 25 to 55, and greater than 55?

- Why is it better to have fewer variables?

Data Exploration – Cross-Tabs

Age Bucket	Retained	Lost	Retained %	Lost %
< 25	869	992	47%	53%
25 - 35	1651	1352	55%	45%
35 - 45	1555	984	61%	39%
45 - 55	1771	847	68%	32%
> 55	1887	592	76%	24%

Clearly, high attrition for customers below the age of 25, and much lower for customers above the age of 55.

- Potential to create age buckets as less than 25, between 25 to 55, and greater than 55?
- Why is it better to have fewer variables?
- Other relevant cross-tabs based on this dataset?

Data Exploration – Cross-Tabs

Table of churn by dur																	
churn	dur																
Frequency Percent	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
0	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	7733 61.86	7733 61.86
1	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	289 2.31	4767 38.14
Total	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	8022 64.18	12500 100.00

Data Exploration – Cross-Tabs

Table of churn by dur																	
churn	dur																
Frequency Percent	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
0	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	7733 61.86	7733 61.86
1	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	289 2.31	4767 38.14
Total	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	8022 64.18	12500 100.00

Data Exploration – Cross-Tabs

Table of churn by dur																	
churn	dur																
Frequency Percent	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
0	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	7733 61.86	7733 61.86
1	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	289 2.31	4767 38.14
Total	1 0.01	1 0.01	68 0.54	227 1.82	238 1.90	217 1.74	273 2.18	251 2.01	212 1.70	196 1.57	209 1.67	933 7.46	639 5.11	367 2.94	329 2.63	317 2.54	8022 64.18	12500 100.00

How would you use this information?

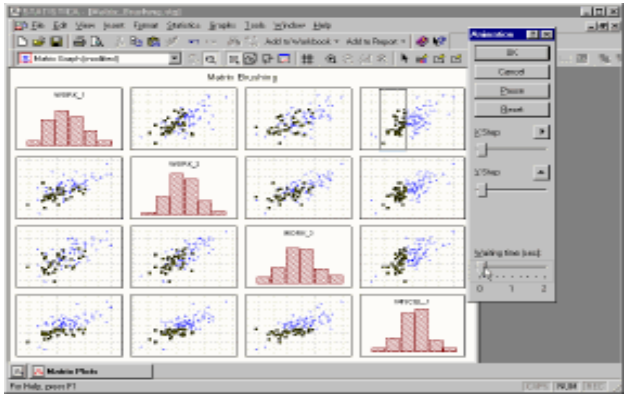
Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative ✓
2. Generate derived variables ✓
3. Summary statistics ✓
4. Cross tabulation ✓
5. Graphical analysis
6. Anomaly detection

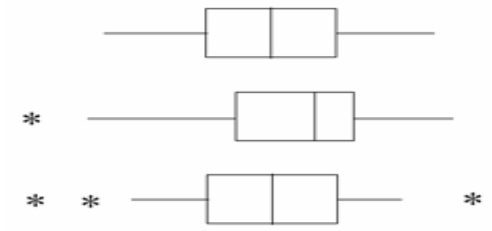
EDA – Graphical Analysis

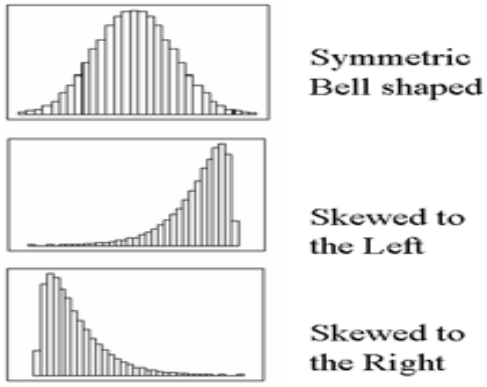
Visualization of data is a powerful method of understanding data and patterns within data



Multiple techniques used to look at data to

- Determine distribution
- Identify spread
- Assess bias / skewness
- Identify outliers





Graphical Data Exploration

1. Useful graphical representations include

Graphical Data Exploration

1. Useful graphical representations include

    I. Simple run charts

Graphical Data Exploration

1. Useful graphical representations include

   I. Simple run charts

   II. Frequency distribution plots
   - Histograms
   - Probability plots

Graphical Data Exploration

1. Useful graphical representations include

   I. Simple run charts

   II. Frequency distribution plots
   - Histograms
   - Probability plots

   III. Range charts
   - Box Plots
   - Stem and Leaf Plots

Graphical Data Exploration

1. Useful graphical representations include

   I. Simple run charts

   II. Frequency distribution plots
   - Histograms
   - Probability plots

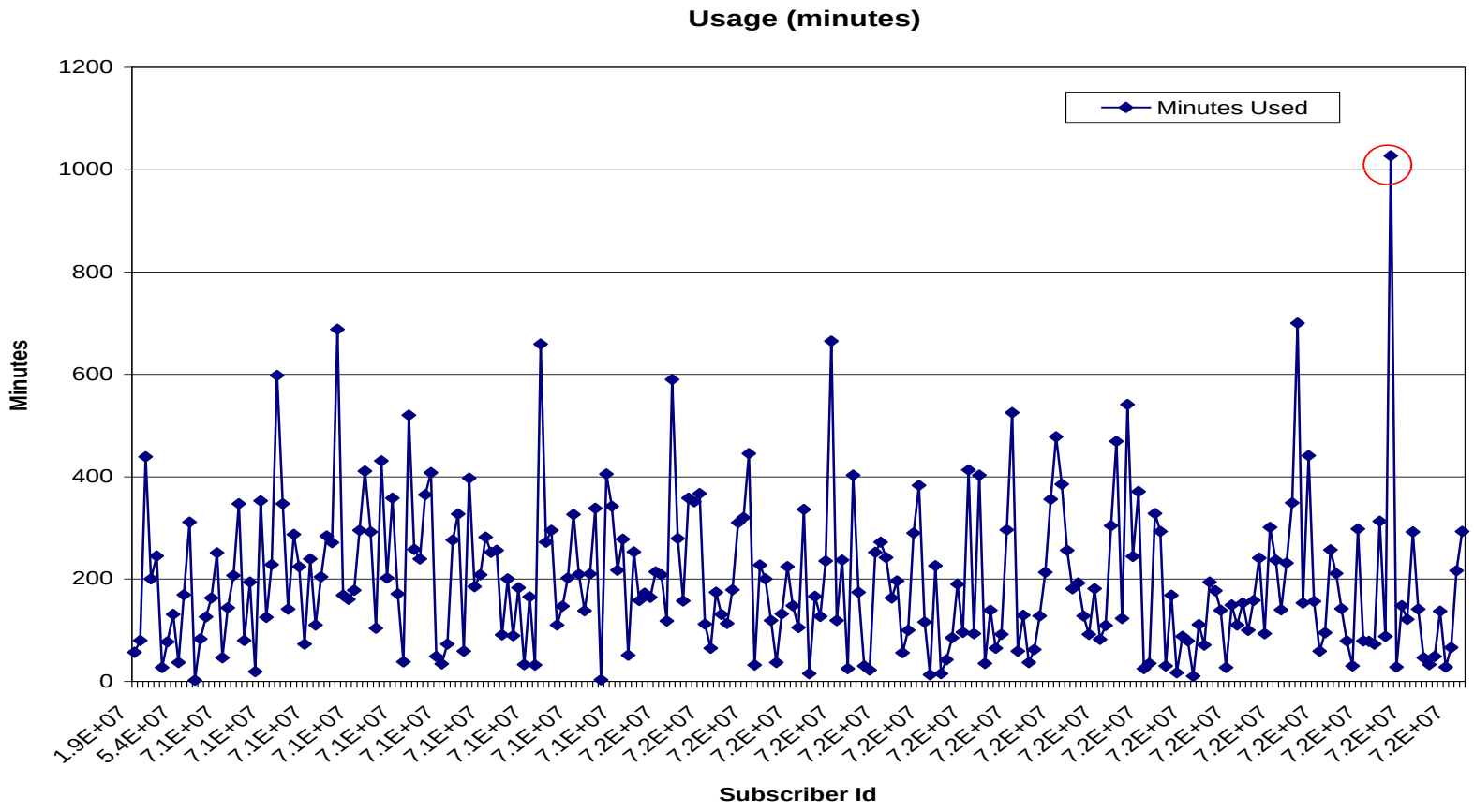
   III. Range charts
   - Box Plots
   - Stem and Leaf Plots

   IV. Joint distribution charts

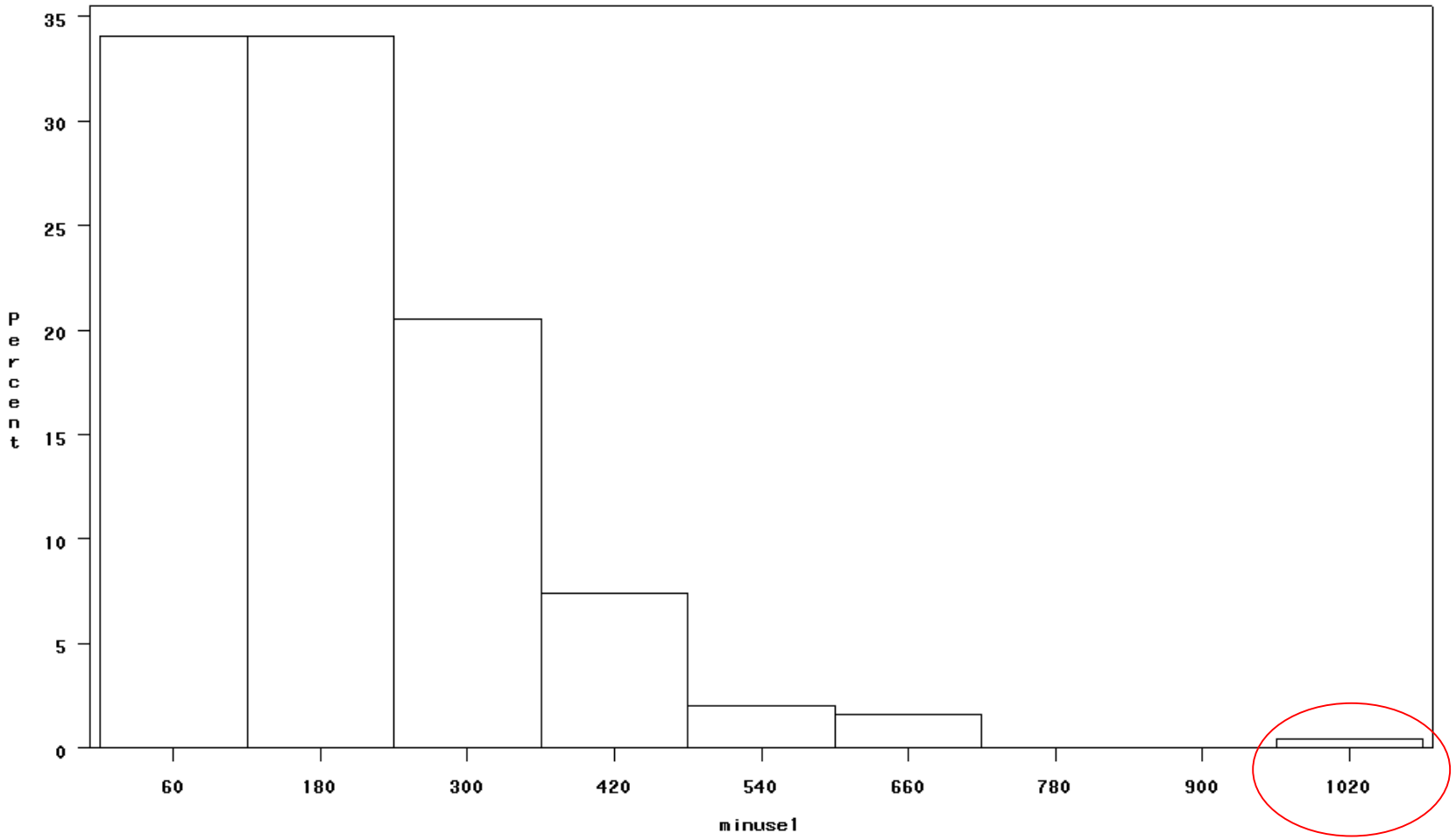
Graphical Data Exploration

Simplest charts for single variables: Run Charts

- Can assess distribution, spread, and outliers



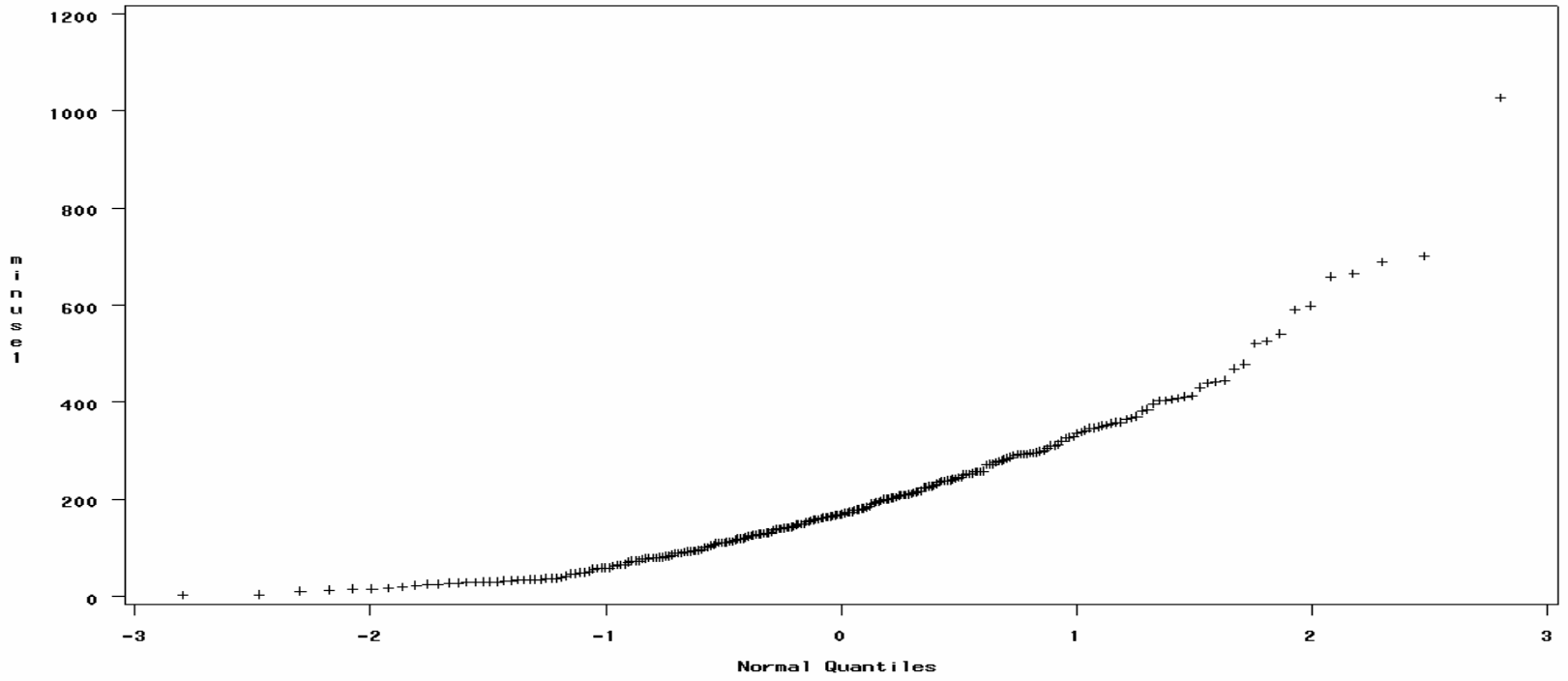
Outlier Identification : Histogram



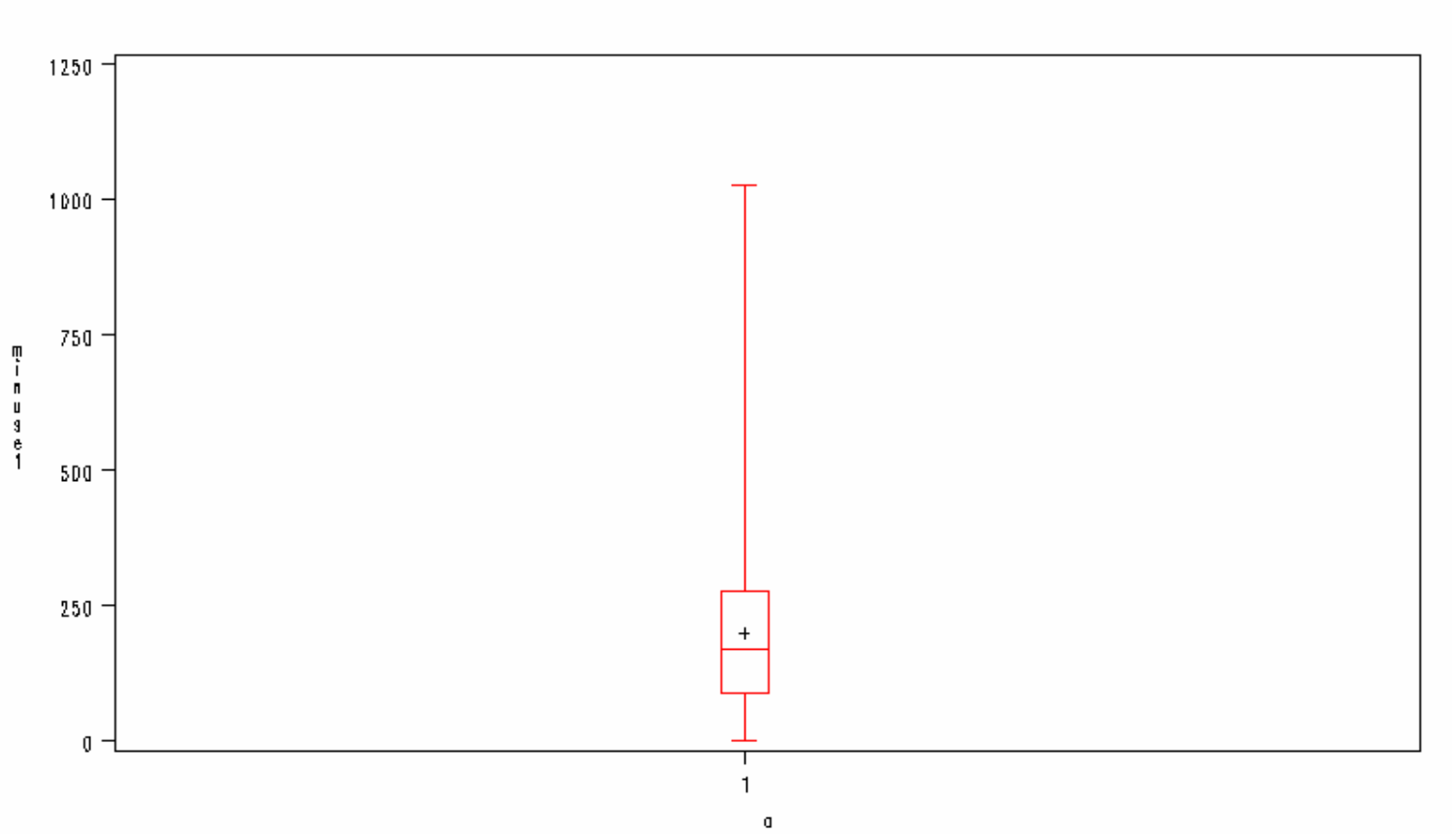
Graphical Data Exploration

Probability Plots – assess distribution shape

- Normal probability plot – widely used to check for normality of distribution

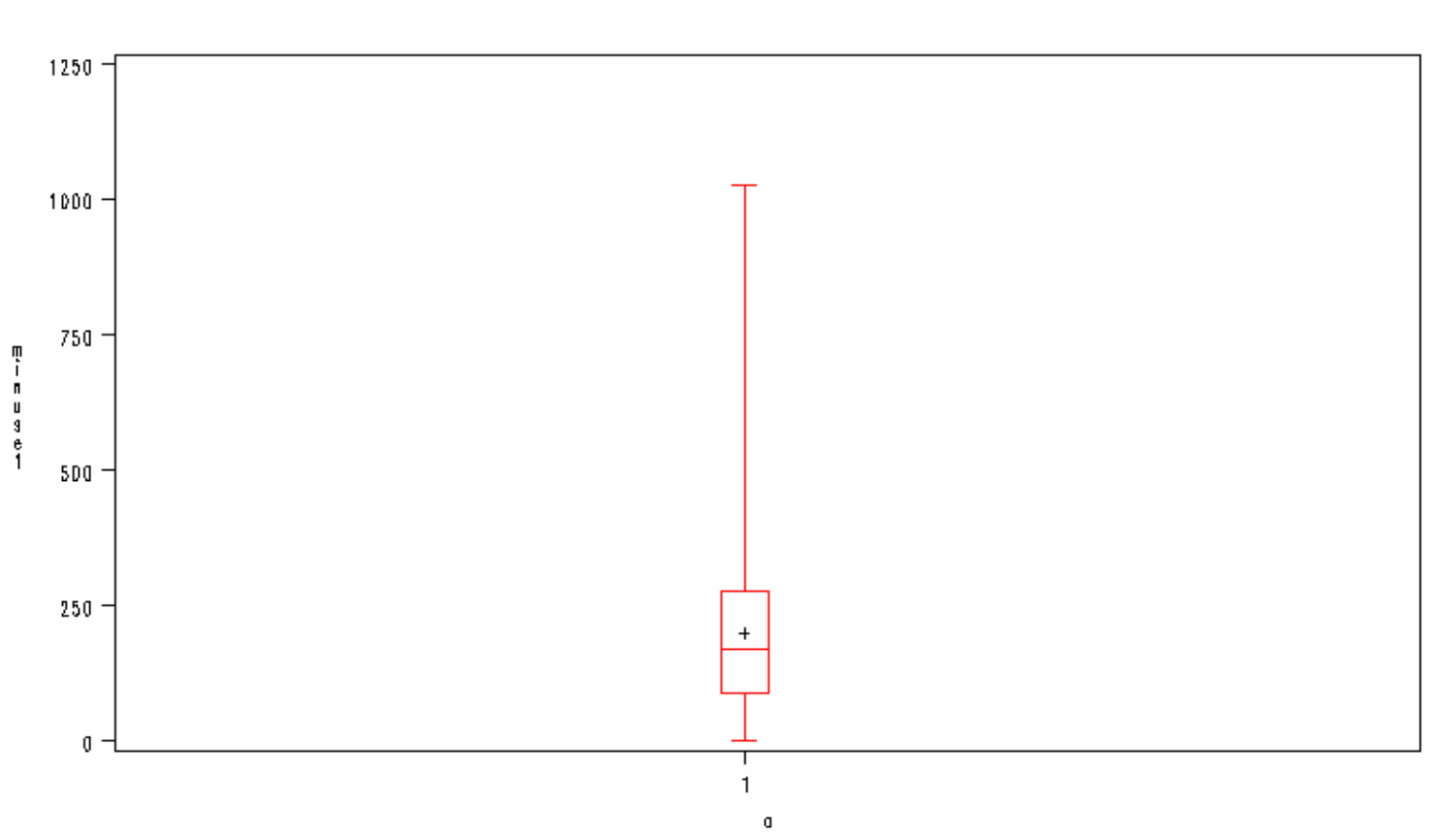


Graphical EDA – Box Plots



- Box plots allow us to look at measures of central tendency explicitly (quartiles are specified)

Graphical EDA – Box Plots



- Box plots allow us to look at measures of central tendency explicitly (quartiles are specified)
- It makes it easier to assess distance for potential outliers

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics
4. Cross tabulation
5. Graphical analysis
6. Anomaly detection

EDA - Exceptions

We have looked at examples of obvious exceptions in previous slides (for example, negative minutes used, etc)

Based on derived variables, need to look at other conditions to identify potential anomalies

- Example: if service end date exists, then do data exist for minutes used post end date?
- Other examples?

EDA - Exceptions

We have looked at examples of obvious exceptions in previous slides (for example, negative minutes used, etc)

Based on derived variables, need to look at other conditions to identify potential anomalies

- Example: if service end date exists, then do data exist for minutes used post end date?
- Other examples?

Exceptions need to be investigated

EDA - Exceptions

We have looked at examples of obvious exceptions in previous slides (for example, negative minutes used, etc)

Based on derived variables, need to look at other conditions to identify potential anomalies

- Example: if service end date exists, then do data exist for minutes used post end date?
- Other examples?

Exceptions need to be investigated

- There may be valid explanations for extreme values

EDA - Exceptions

We have looked at examples of obvious exceptions in previous slides (for example, negative minutes used, etc)

Based on derived variables, need to look at other conditions to identify potential anomalies

- Example: if service end date exists, then do data exist for minutes used post end date?
- Other examples?

Exceptions need to be investigated

- There may be valid explanations for extreme values
- Danger of making data set very “general”

What is Data Exploration?

1. What is the information contained in the data?
2. What is the quality of this information?
3. **Is the data complete?**

EDA Summary

- It is very critical to invest time in understanding the data once it has been pulled/received, since it is really the starting point to building any solution

EDA Summary

- It is very critical to invest time in understanding the data once it has been pulled/received, since it is really the starting point to building any solution

- Model results will only be as good as the data that goes in; and model results be reliable only if the right models are applied to the existing data

EDA Summary

- It is very critical to invest time in understanding the data once it has been pulled/received, since it is really the starting point to building any solution
- Model results will only be as good as the data that goes in; and model results be reliable only if the right models are applied to the existing data
- There are many techniques of EDA that can be used to gain a real understanding of available data, relationships between variables, and potential issues with available data

Recap – Data Exploration

Data Assessment:

1. What is the information contained in the data?
2. What is the quality of this information?
3. Is the data complete?

Assessing Data Quality

6 steps in data exploration

1. Transform qualitative into quantitative
2. Generate derived variables
3. Summary statistics
4. Cross tabulation
5. Graphical analysis
6. Anomaly detection

EDA Output

EDA Output

Internal audience / analyst:

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables

2. Identification and assessment of exceptions / outliers / wrong values / missing values

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables

2. Identification and assessment of exceptions / outliers / wrong values / missing values

3. Follow up questions

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables
2. Identification and assessment of exceptions / outliers / wrong values / missing values
3. Follow up questions

External audience:

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables
2. Identification and assessment of exceptions / outliers / wrong values / missing values
3. Follow up questions

External audience:

1. Presentation summarizing initial data assessment

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables
2. Identification and assessment of exceptions / outliers / wrong values / missing values
3. Follow up questions

External audience:

1. Presentation summarizing initial data assessment
2. Visuals showing interesting/useful patterns in dataset

EDA Output

Internal audience / analyst:

1. Understanding of data and all variables
2. Identification and assessment of exceptions / outliers / wrong values / missing values
3. Follow up questions

External audience:

1. Presentation summarizing initial data assessment
2. Visuals showing interesting/useful patterns in dataset
3. Follow up questions

EDA Output

Internal audience / analyst:

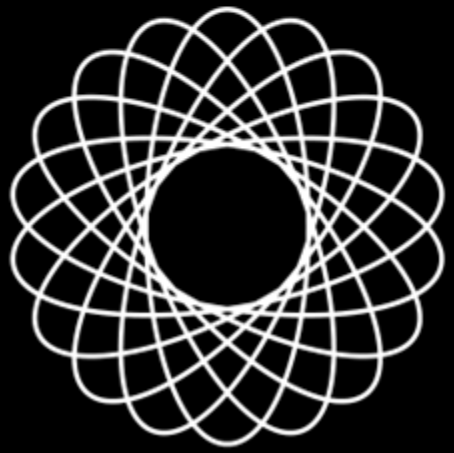
1. Understanding of data and all variables
2. Identification and assessment of exceptions / outliers / wrong values / missing values
3. Follow up questions

External audience:

1. Presentation summarizing initial data assessment
2. Visuals showing interesting/useful patterns in dataset
3. Follow up questions
4. Next steps

THANK YOU

DATA SCIENCE



DATA SCIENCE

★ Continued Lab Sessions ★

Data Manipulation: DATA EXPLORATION

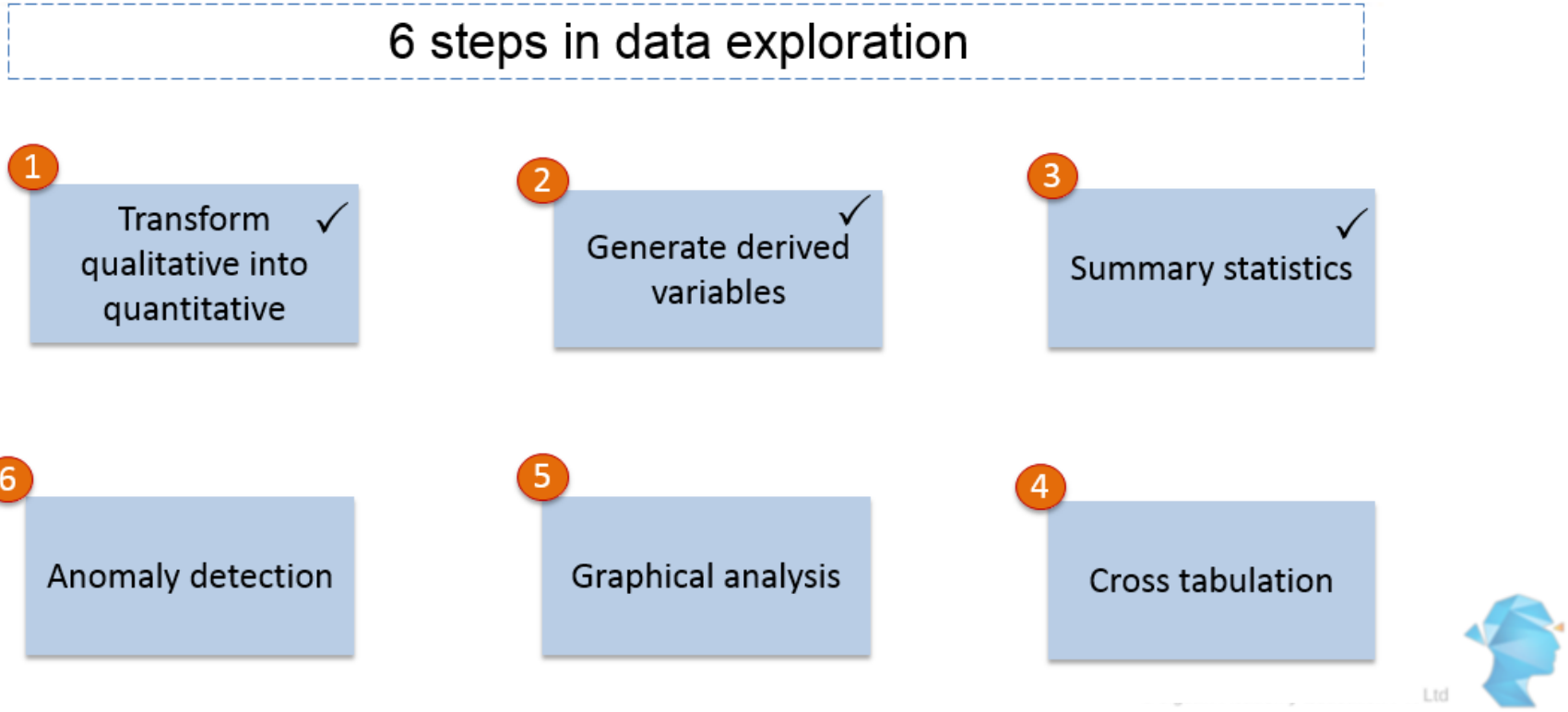
DATA EXPLORATION

Type of variable	Variable	Description	Data Type	Data Values
Demographic Variables	Age	Age of customer as of date of data collection	Numeric	Years
	Marital Status	Marital status of customer	Character, Categorical	"Married", "Divorced", "Single"
	Job Type	Type of job	Character, Categorical	"admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services"
	Education Level	Education level completed	Character, Categorical	"unknown","secondary","primary","tertiary"
Financial variables	Credit default	Does the customer have some loans in default	Character, Binary	"yes","no"
	Balance	Average Yearly Balance (in Euros)	Numeric	Euros
	Housing Loan	Does the customer have an existing housing loan?	Character, Binary	"yes","no"
	Personal Loan	Does the customer have oan existing personal loan?	Character, Binary	"yes","no"
Telemarketing history	Contact communication type	Type of communication used in the last contact	Character, Categorical	unknown,"telephone","cellular"
	Day of last contact	last contact day of the month	Numeric	1-31
	Month of last contact	last contact month of year	Numeric	"jan", "feb", "mar", ..., "nov", "dec"
	Duration	last contact duration, in seconds	Numeric	Seconds
	Current Campaign Contacts	number of contacts performed during this campaign and for this client	Numeric, includes last contact	
	Days since last contact	number of days that passed by after the client was last contacted from a previous campaign	Numeric	-1 means client was not previously contacted
	Previous Contacts	number of contacts performed before this campaign and for this client	Numeric	Number
	Previous Outcome	outcome of the previous marketing campaign	Character, Categorical	"unknown,"other","failure","success"
Target Variable	Outcome	has the client subscribed a term deposit?	Character, Binary	"yes","no"

DATA EXPLORATION

The aim of data exploration is to:

1. Assess the data quality and completeness
2. Get a big picture understanding of individual variables and overall trends
3. Identify any potential data issues and ways to manage them



DATA EXPLORATION

Data Exploration can be performed using Excel, SAS or a combination of both

Number	Variable	Type
9	contact	Char
5	default	Char
4	education	Char
7	housing	Char
2	job	Char
8	loan	Char
3	marital	Char
11	month	Char
16	poutcome	Char
17	y	Char
1	age	Num
6	balance	Num
13	campaign	Num
10	day	Num
12	duration	Num
14	pdays	Num
15	previous	Num

DATA EXPLORATION

The first step is to generate a high level understanding of what is contained in the data

Variable	Type	Description	Values	Distribution	Distribution				Comments
contact	Char	Type of communication used in the last contact	Categorical	Cellular: 64.3% telephone: 6.5% unknown: 28.8%	Cellular: 64.3%	Telephone : 6.5%	Unknown: 28.8%		30% missing
default	Char	Does the customer have some loans in default	Categorical	no: 98.2% yes: 1.8%	No: 98.2%	Yes 1.8%			Cross Tab with Y variable to see if useful?
education	Char	Education level completed	Categorical	primary: 15.15% secondary: 51.3% tertiary:29.4% unknown: 4.1%	Primary: 15.15%	Secondary : 51.3%	Tertiary: 29.4%	Unknown: 4.1%	
housing	Char	Does the customer have an existing housing loan?	Categorical	no: 44.4% yes: 55.6%	no: 44.4%	yes 55.6%			
job	Char	Type of job	Categorical	admin: 11.4%, blue collar: 21.5%	bluecollar: 21.5%	managem ent: 20.9%	services: 9.2%	technician 16.8%	Combine entrepenuer with mgmt, self employed, housemaid with services, and retited, unemployed, student as no

DATA EXPLORATION

Descriptive summary statistics for continuous data

Variable	Minimum	Maximum	Mean	nmiss	Median	Std
age	18	95	41	0	39	11
balance	-8019	102127	1362	0	448	3045
day	1	31	16	0	16	8
duration	0	4918	258	0	180	258
campaign	1	63	3	0	2	3
pdays	-1	871	40	0	-1	100
previous	0	275	1	0	0	2

Negative balance - wrong data? Check incidence of negatives

Balance: Max is an outlier?

Duration: check max for outliers?

Campaign: check max for outliers

pdays: -1 is no contact. Remove and make 0

previous: check max for outlier

DATA EXPLORATION

Distribution of the balance variable shows multiple negative values

Possibly overdraft facilities?

The maximum value is much higher than 99% value,
but looking at top 5 values shows a gradual increase.

May want to exclude high values, but first, check if behaviour is different for high values:

# of obs where balance > 20000 = 193
Proportion of No responses in these obs: 85%

Negative Balance:

Max

Quantiles (definition 5)	
Quantile	Estimate
100% Max	102127
99%	13165
95%	5768
90%	3574
75% Q3	1428
50% Median	448
25% Q1	72
10%	0
5%	-172
1%	-627
0% Min	-8019

Extreme Observations			
Value	Obs	Value	Obs
-8019	12910	71188	41694
-6847	15683	81204	42559
-4057	38737	81204	43394
-3372	7414	98417	26228
-3313	1897	102127	39990

Many negative observations - OD?
High value may
not be an

# of obs > 20000:

y	Frequency	Percent
no	164	84.97
ye	29	15.03

193 obs
% of yes/no not different from total sample

DATA EXPLORATION

Distribution of the balance variable shows multiple negative values

Possibly overdraft facilities?

The maximum value is much higher than 99% value,
but looking at top 5 values shows a gradual increase.

May want to exclude high values, but first, check if behaviour is different for high values:

# of obs where balance > 20000 = 193
Proportion of No responses in these obs: 85%

Negative Balance:

Max

Quantiles (definition 5)	
Quantile	Estimate
100% Max	102127
99%	13165
95%	5768
90%	3574
75% Q3	1428
50% Median	448
25% Q1	72
10%	0
5%	-172
1%	-627
0% Min	-8019

Extreme Observations			
Value	Obs	Value	Obs
-8019	12910	71188	41694
-6847	15683	81204	42559
-4057	38737	81204	43394
-3372	7414	98417	26228
-3313	1897	102127	39990

Many negative observations – OD?
High value may
not be an

# of obs > 20000:

y	Frequency	Percent
no	164	84.97
ye	29	15.03

193 obs
% of yes/no not different from total sample

DATA EXPLORATION

Count of y	Column Labels			
Row Labels	no	ye	(blank)	Grand Total
failure	4283	618		4901
other	1533	307		1840
success	533	978		1511
unknown	33573	3386		36959
(blank)				
Grand Total	**39922**	**5289**		**45211**

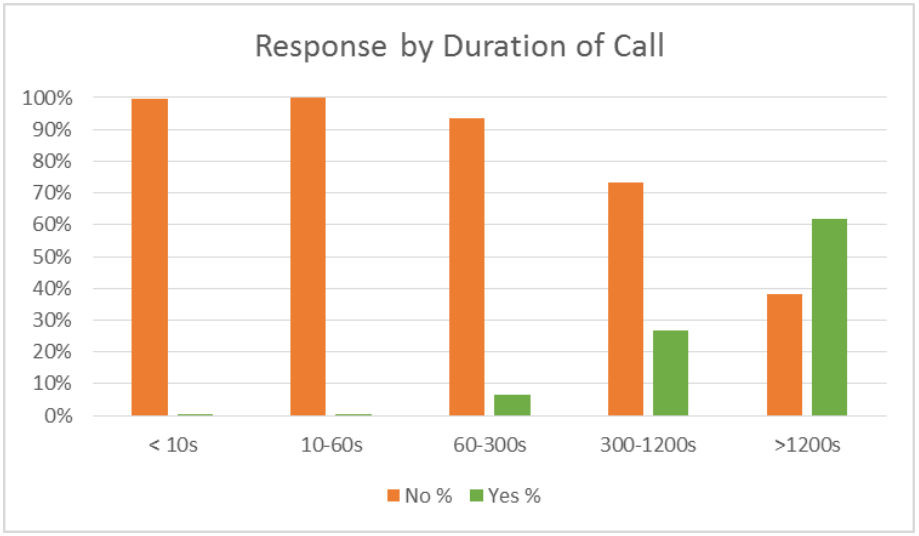
Count of y	Column Labels		
Row Labels	no	ye	Grand Total
no	16727	3354	20081
yes	23195	1935	25130
(blank)			
Grand Total	**39922**	**5289**	**45211**

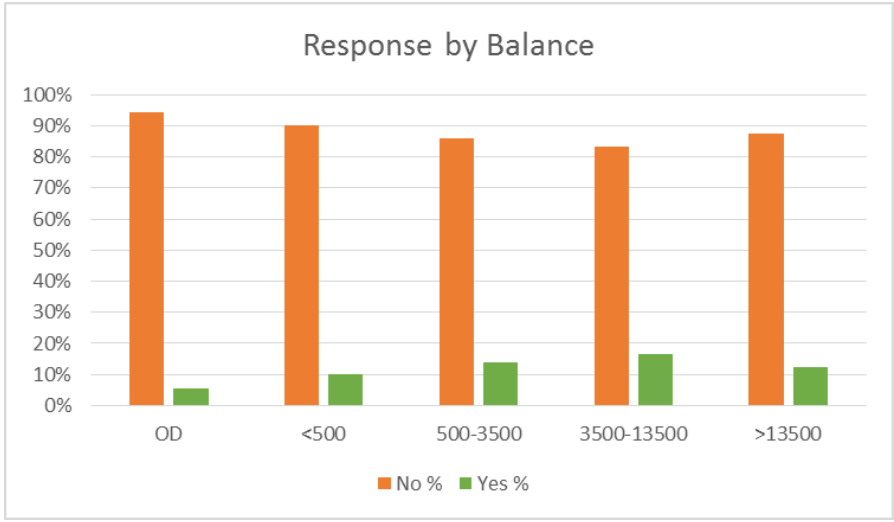
DATA EXPLORATION

Count of y	Column Labels			
Row Labels	no	ye	(blank)	Grand Total
failure	4283	618		4901
other	1533	307		1840
success	533	978		1511
unknown	33573	3386		36959
(blank)				
Grand Total	**39922**	**5289**		**45211**

Count of y	Column Labels		
Row Labels	no	ye	Grand Total
no	16727	3354	20081
yes	23195	1935	25130
(blank)			
Grand Total	**39922**	**5289**	**45211**

DATA EXPLORATION





DATA EXPLORATION - FINDINGS

Missing Data:

Previous Outcome: 81% missing – cannot be used in analysis

Contact type: 29% missing – recommend dropping from analysis

Outliers:

One outlier in Previous (value 275, next highest value 16)

Potential outliers:

- balance: > 20k
- duration > 2000
- pdays > 550
- campaign > 20

DATA EXPLORATION - FINDINGS

Other findings:

Overall response percent in the dataset: 15%

Bivariate data analysis suggests that age, call duration, job type are potentially strong influencers of response

Previous outcome looks to have a strong impact on outcome, but 80% data is missing