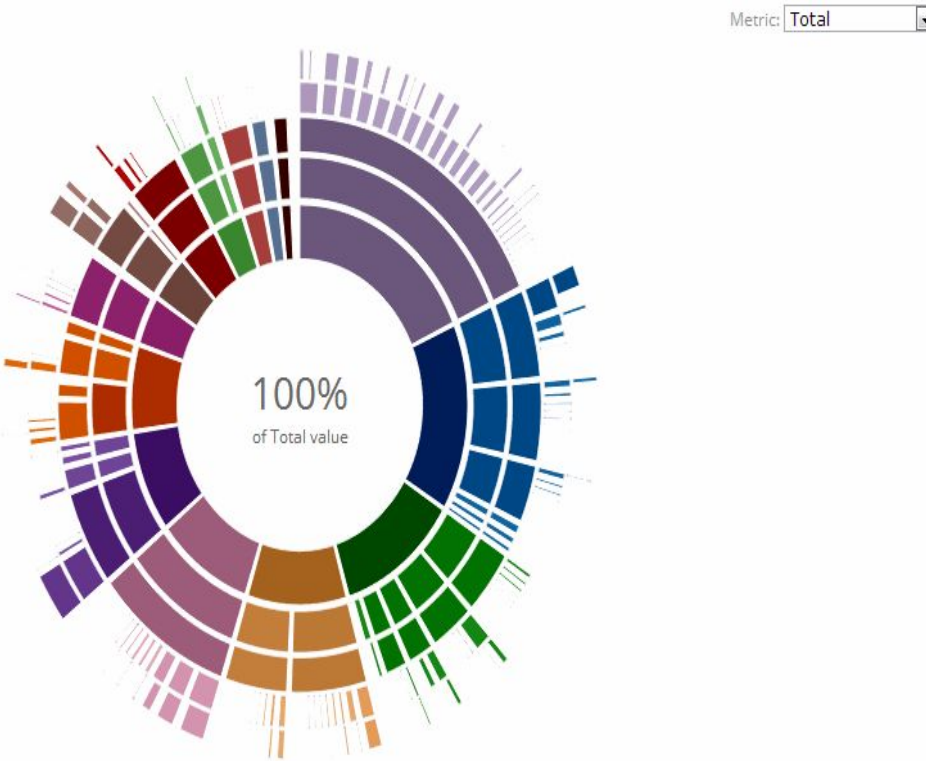
Data Analytics

Professor Ernesto Lee

Business Analytics

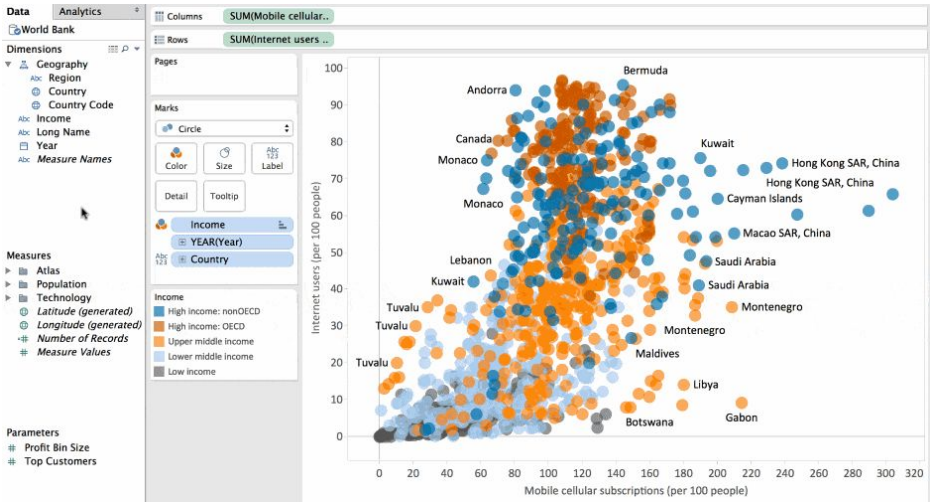
“Be approximately right rather than exactly wrong”
- John Turkey



Business Problem

Please carry out an Exploratory Data Analysis and create a compelling story based on the given dataset; also predict which Article will be more popular in the near future.

Tools



1. R Studio or Python
2. SPSS
3. Power BI
4. Excel
5. Orange
6. Tableau

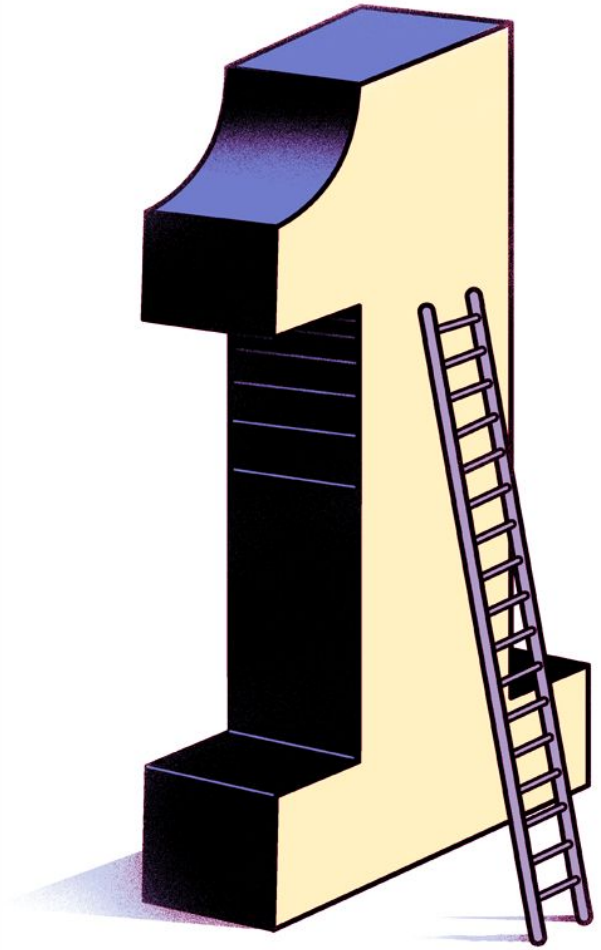
Exploratory Data Analysis

1. Variable Identification
2. Univariate Analysis
3. Bi-variate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation
7. Variable creation

Stage One of Your Analysis

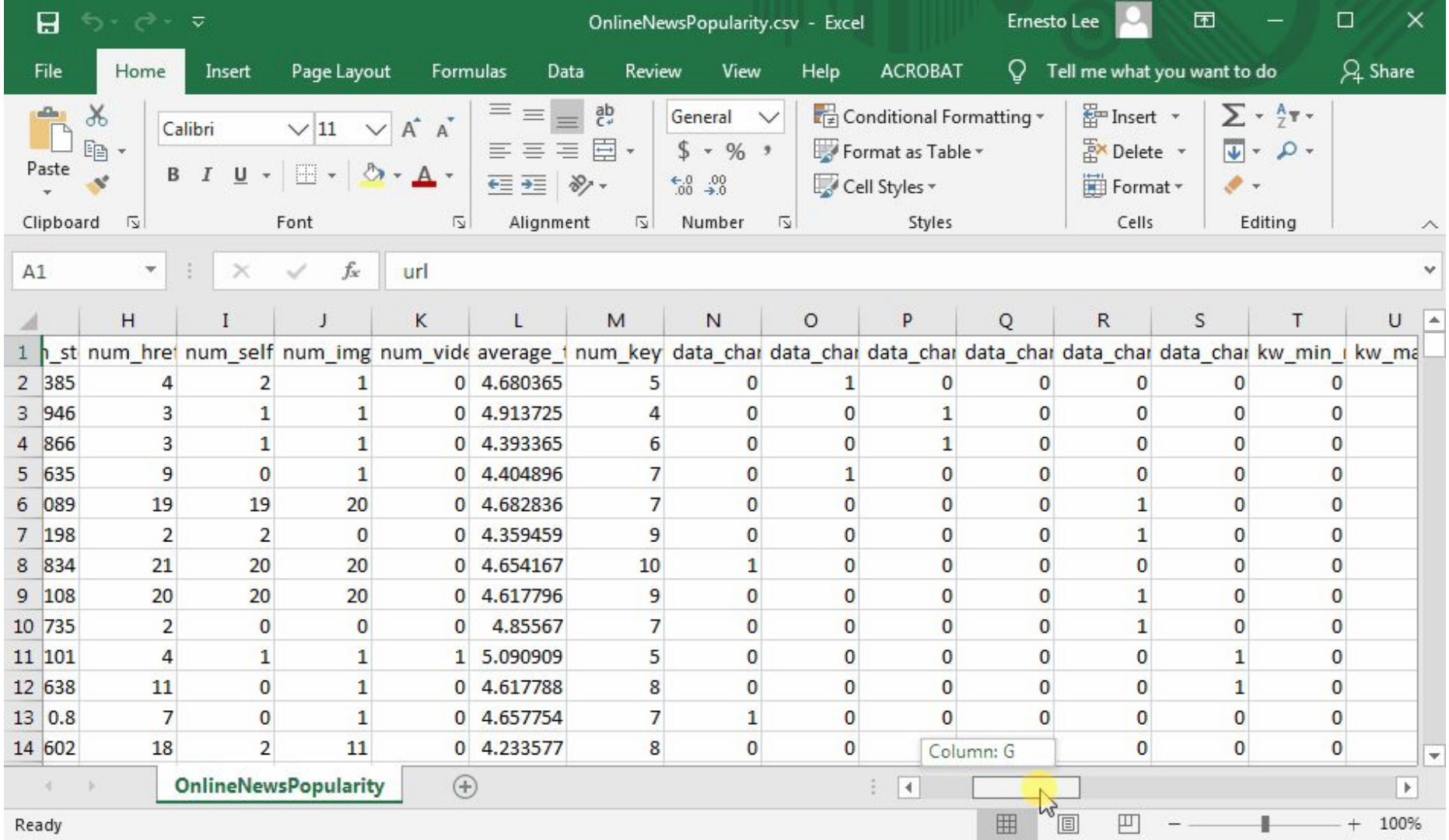
Stage One of Your Analysis

1. **Identify your TARGET VARIABLE from your dataset. (What are you trying to analyze or predict)**
2. **Identify all DATA TYPES (Continuous, Discrete, Ordinal, Categorical, etc.)**
3. **Identify which columns you actually need.**



What is Your Target?

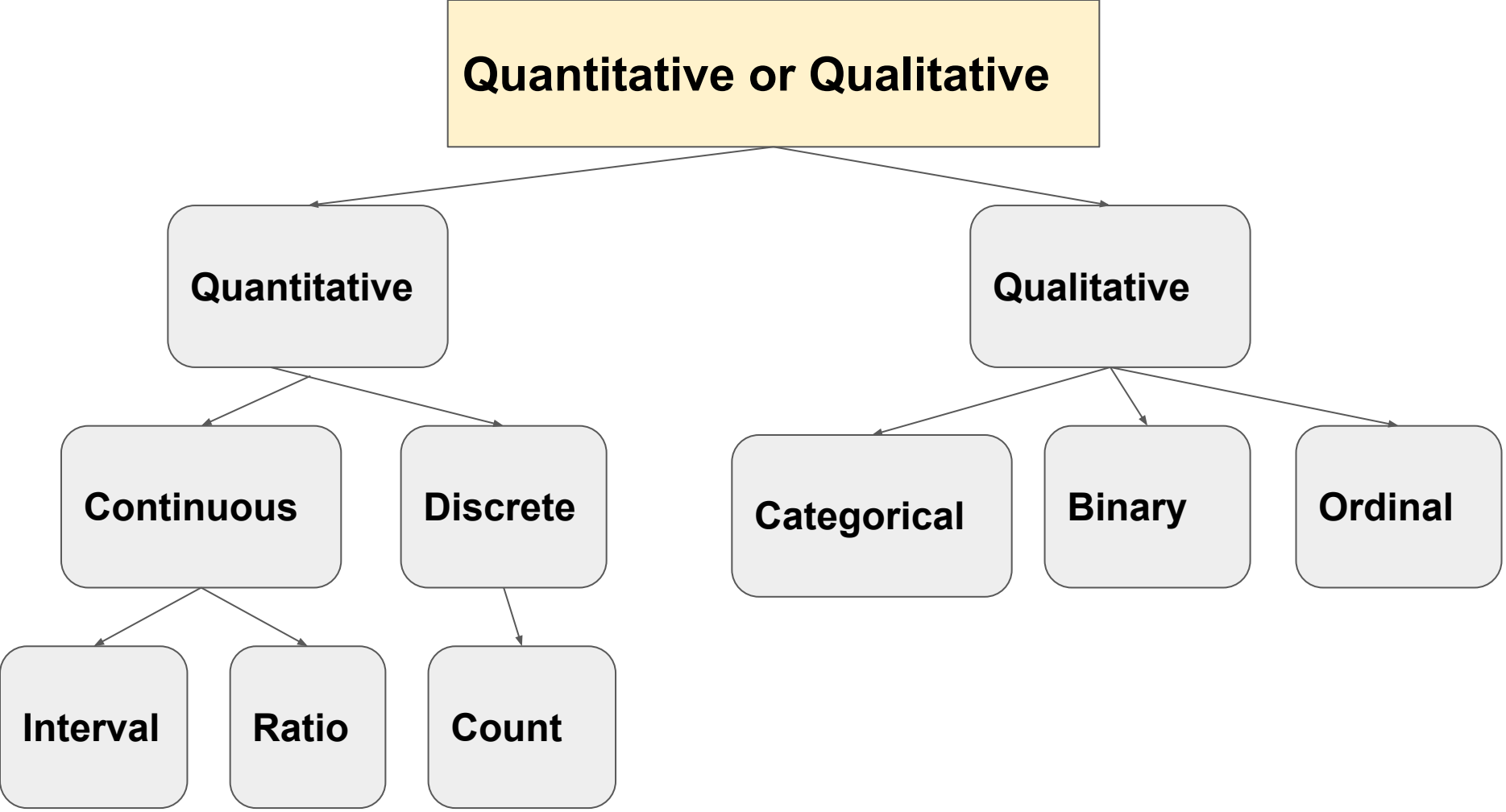
Look at your dataset and determine which feature (column) of data gives you the best chance of answering your business question.

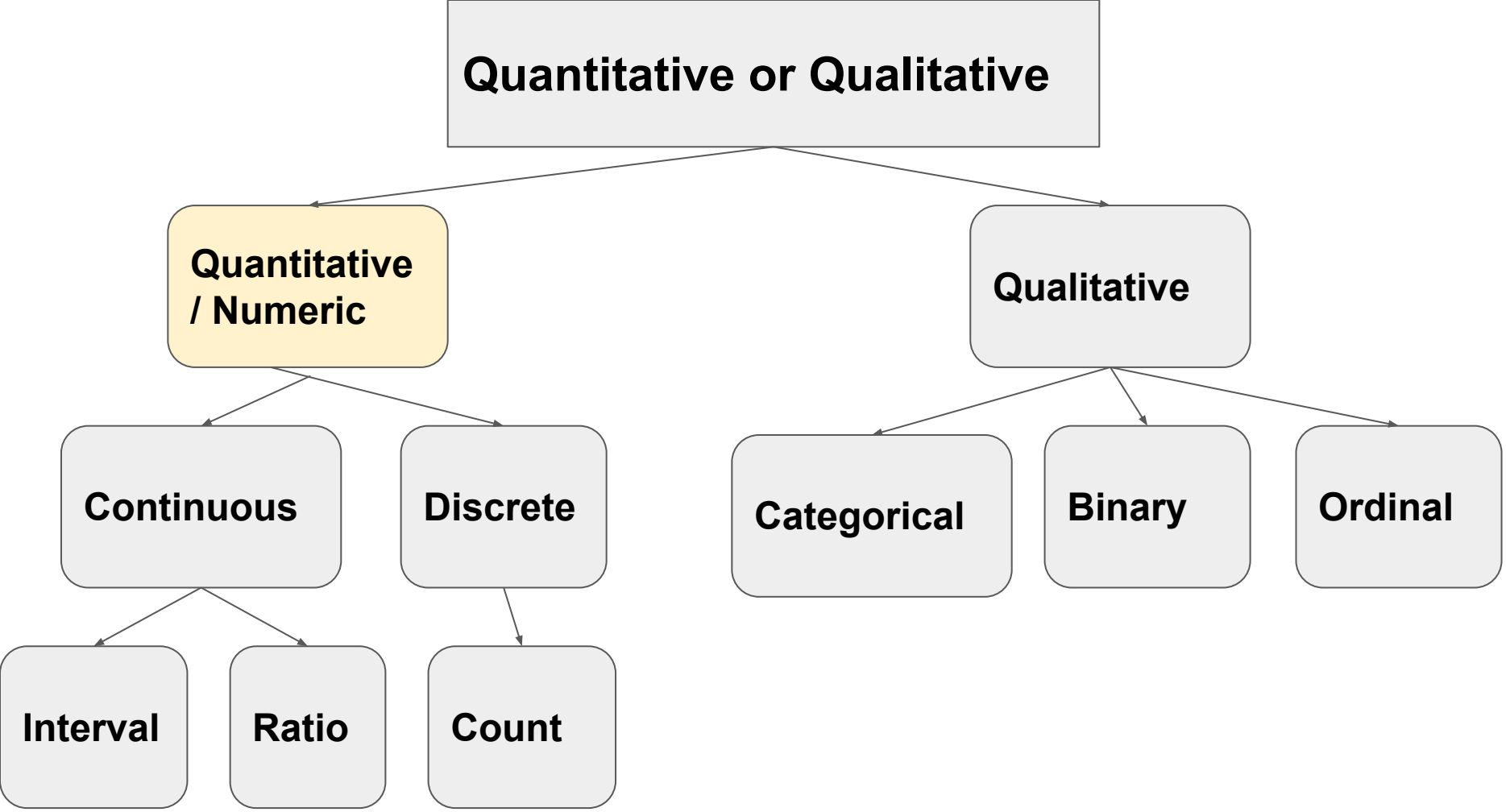


	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	h_st	num_hre	num_self	num_img	num_vide	average_t	num_key	data_char	data_char	data_char	data_char	data_char	data_char	kw_min_	kw_ma
2	385	4	2	1	0	4.680365	5	0	1	0	0	0	0	0	
3	946	3	1	1	0	4.913725	4	0	0	1	0	0	0	0	
4	866	3	1	1	0	4.393365	6	0	0	1	0	0	0	0	
5	635	9	0	1	0	4.404896	7	0	1	0	0	0	0	0	
6	089	19	19	20	0	4.682836	7	0	0	0	0	1	0	0	
7	198	2	2	0	0	4.359459	9	0	0	0	0	1	0	0	
8	834	21	20	20	0	4.654167	10	1	0	0	0	0	0	0	
9	108	20	20	20	0	4.617796	9	0	0	0	0	1	0	0	
10	735	2	0	0	0	4.85567	7	0	0	0	0	1	0	0	
11	101	4	1	1	1	5.090909	5	0	0	0	0	0	1	0	
12	638	11	0	1	0	4.617788	8	0	0	0	0	0	1	0	
13	0.8	7	0	1	0	4.657754	7	1	0	0	0	0	0	0	
14	602	18	2	11	0	4.233577	8	0	0			0	0	0	

Identify the Data Types for Every Feature (Column)

At this point - you have to identify every data type for every column that you are interested in.


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative / Numeric
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count

Warning about Quantitative versus Qualitative Data

If the data is numeric BUT it is “encoded” to represent a thing, then it is NOT numeric - it is QUALITATIVE even though it is represented by NUMBERS.

Broward College Central Campus - 1

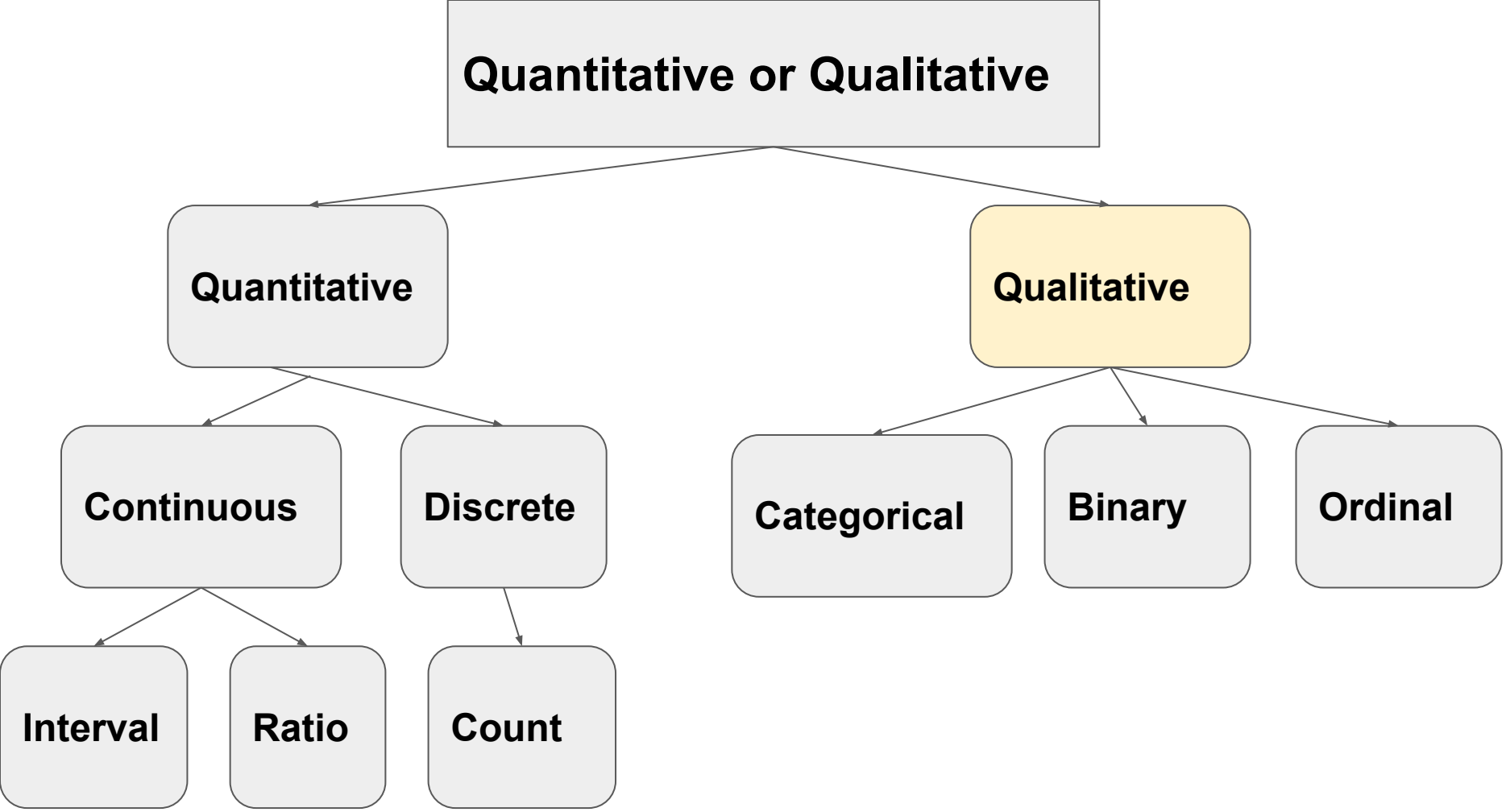
North Campus = 2

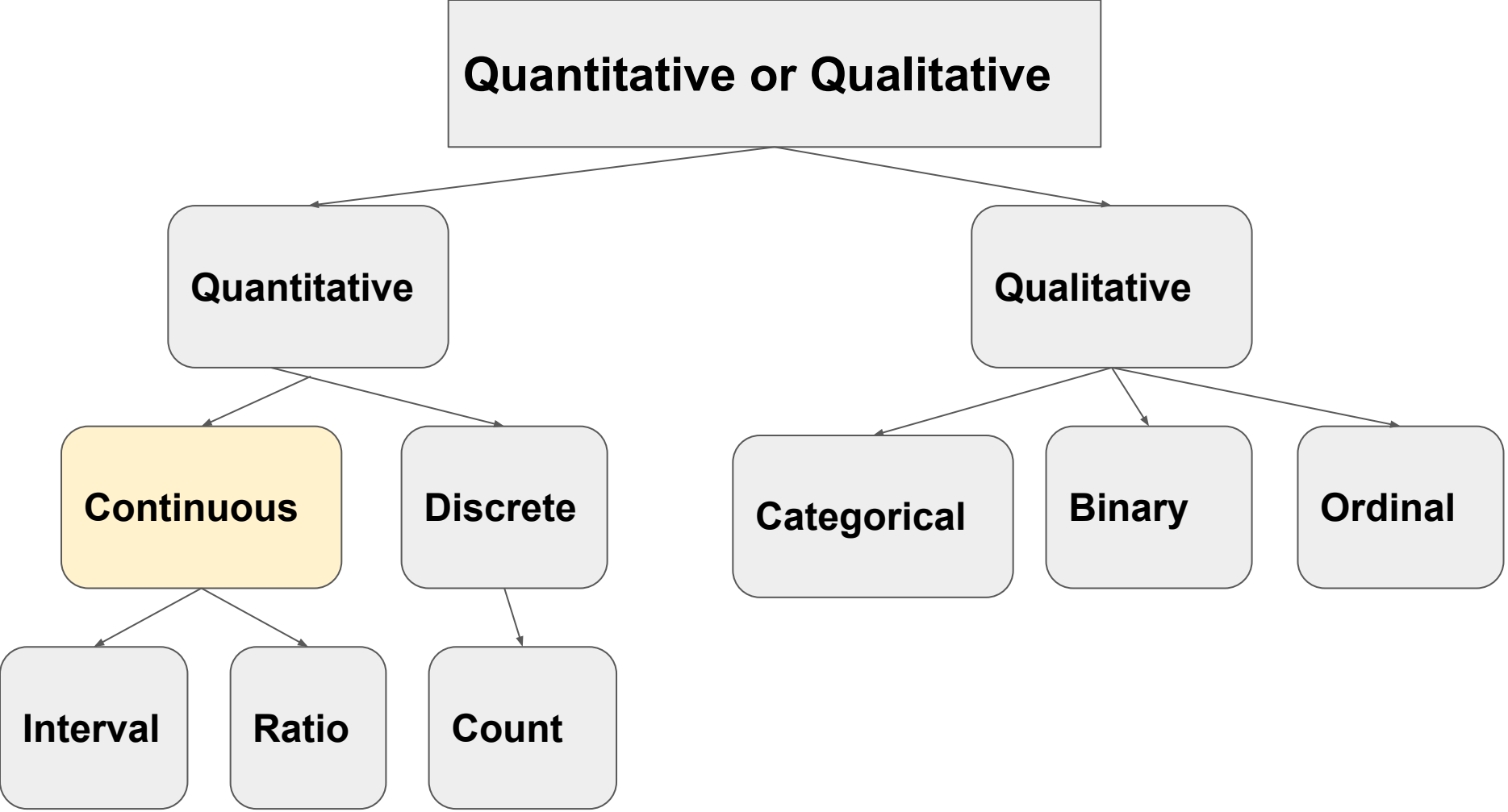
South Campus = 3

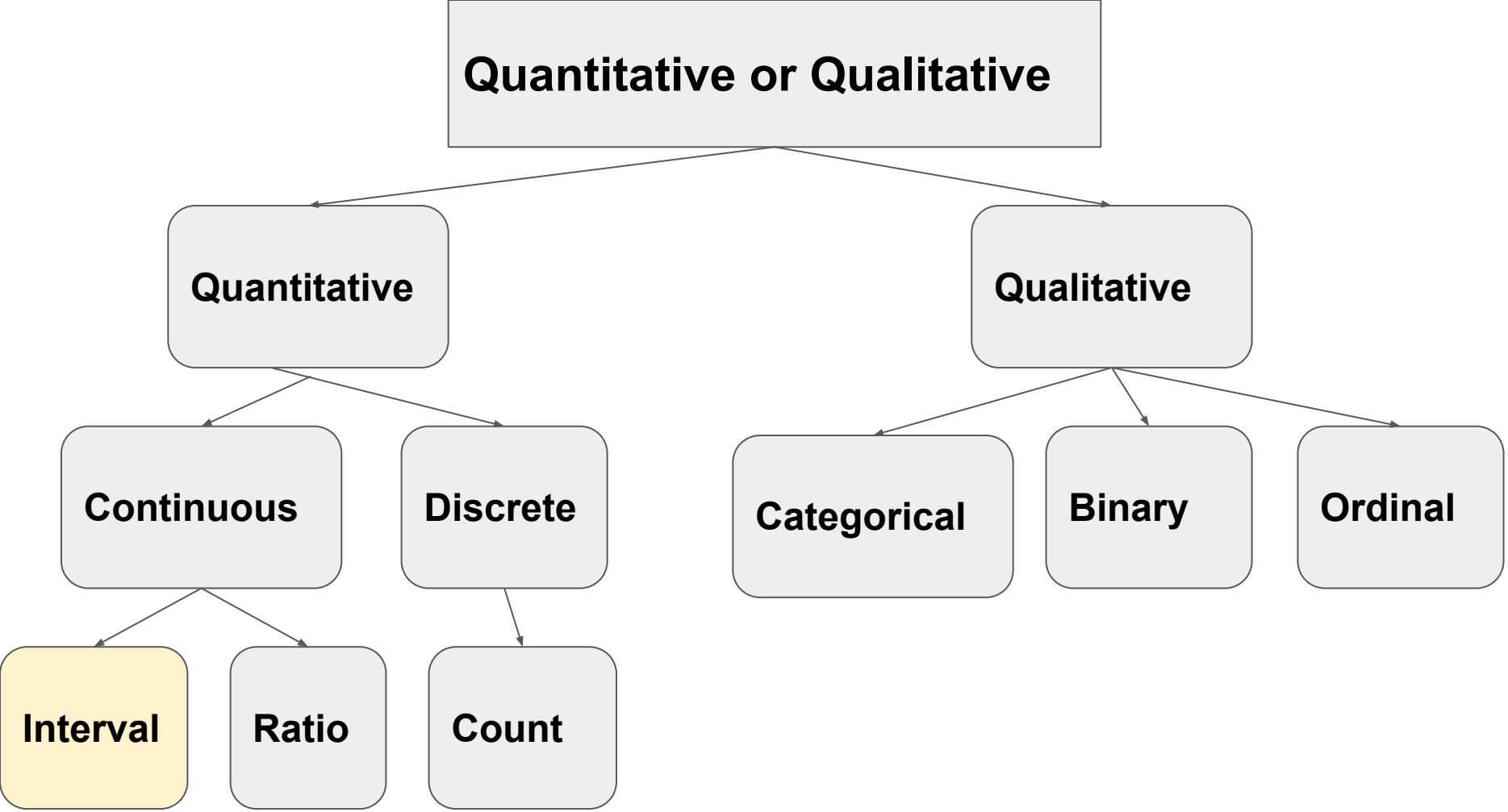
Downtown = 4

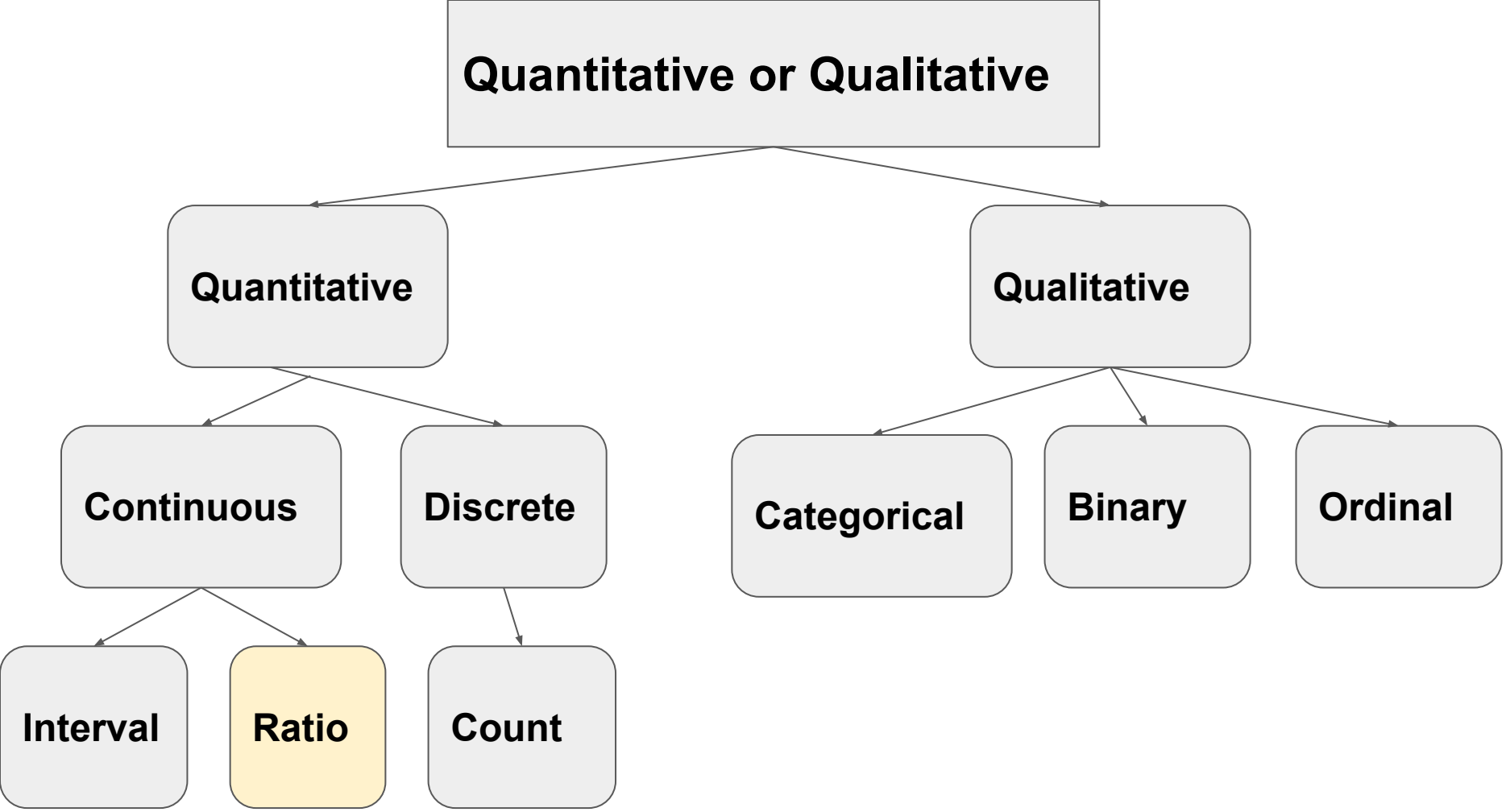
Online = 5

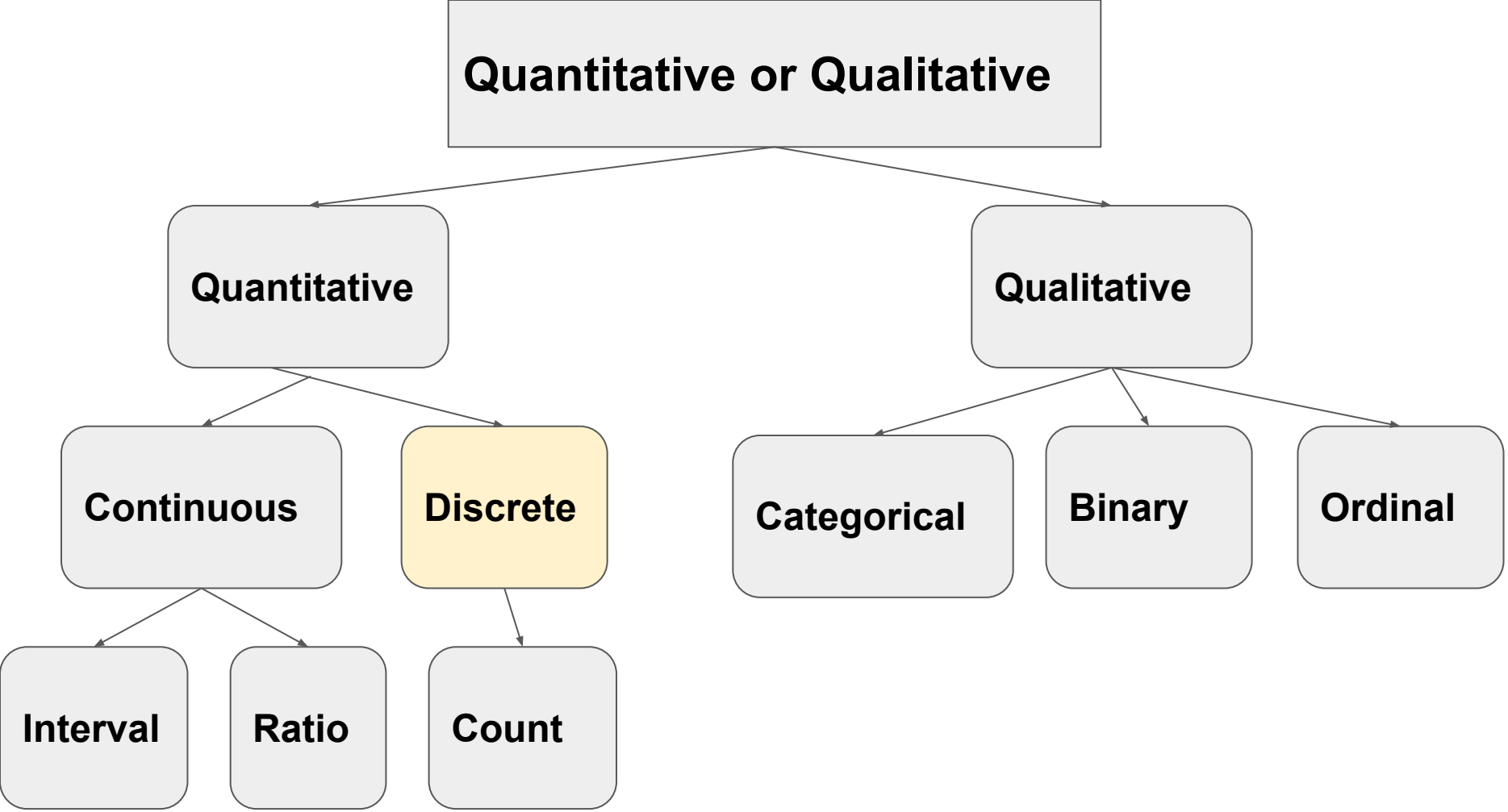
etc.

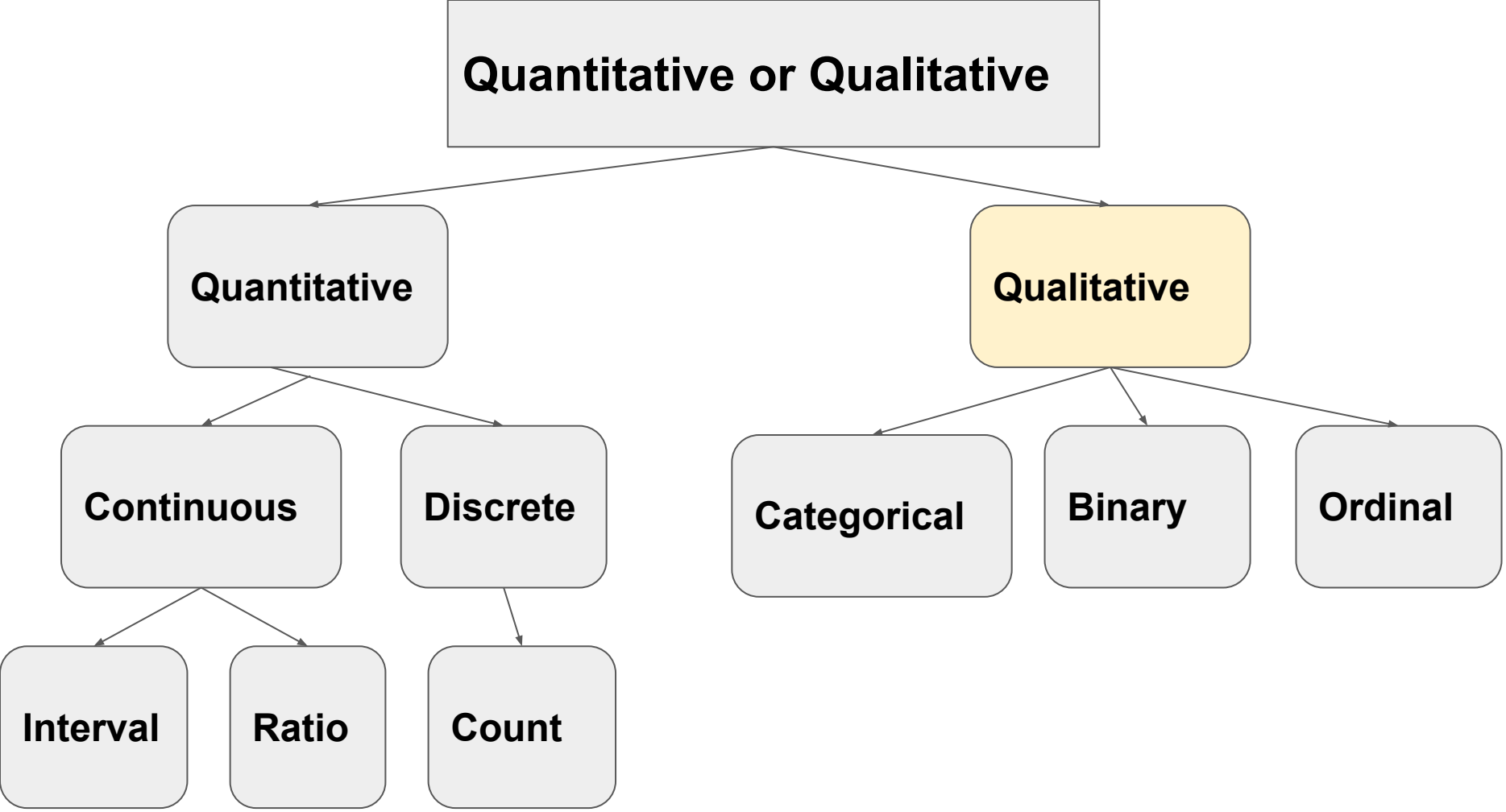

Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count

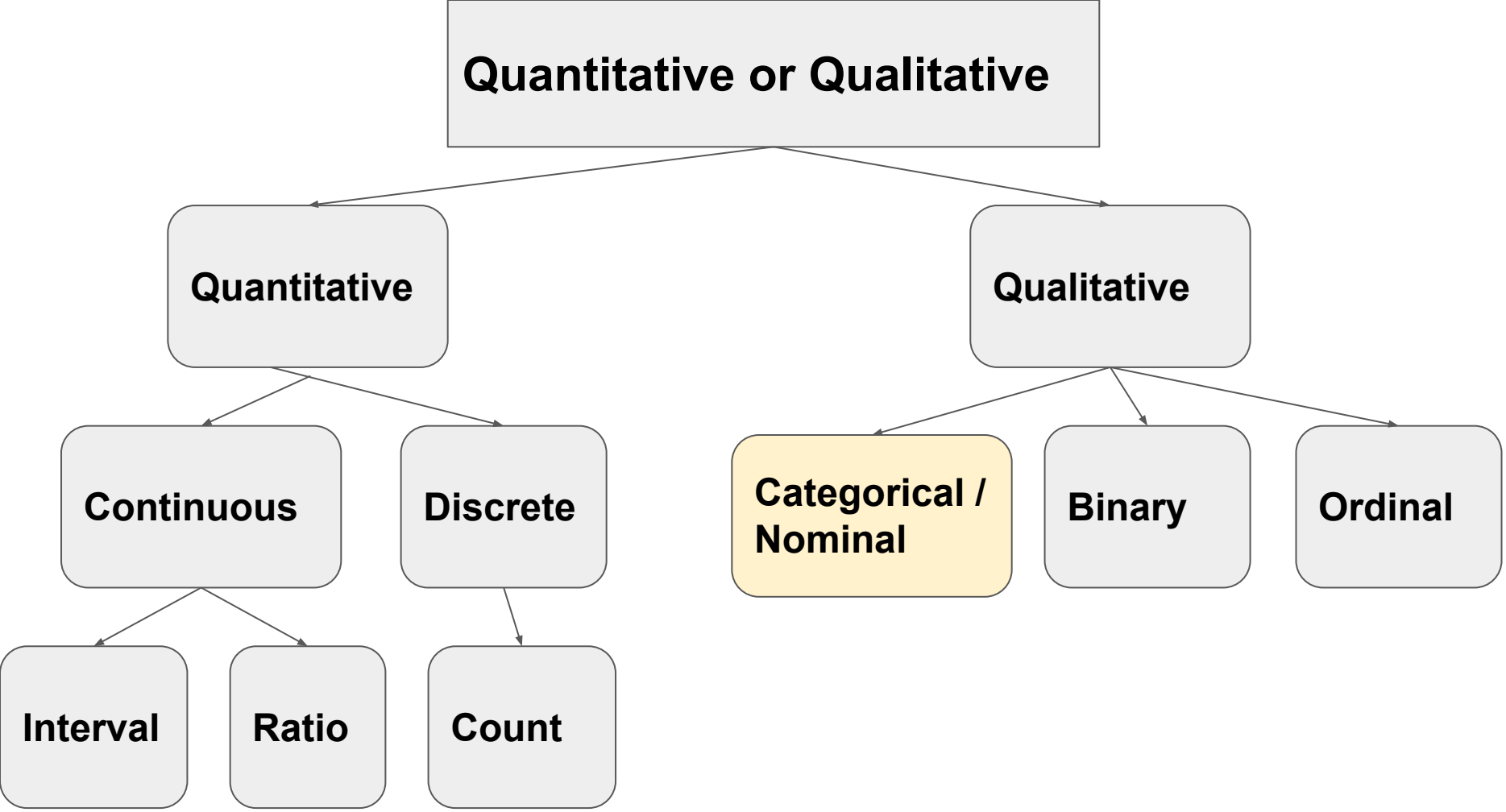

Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count

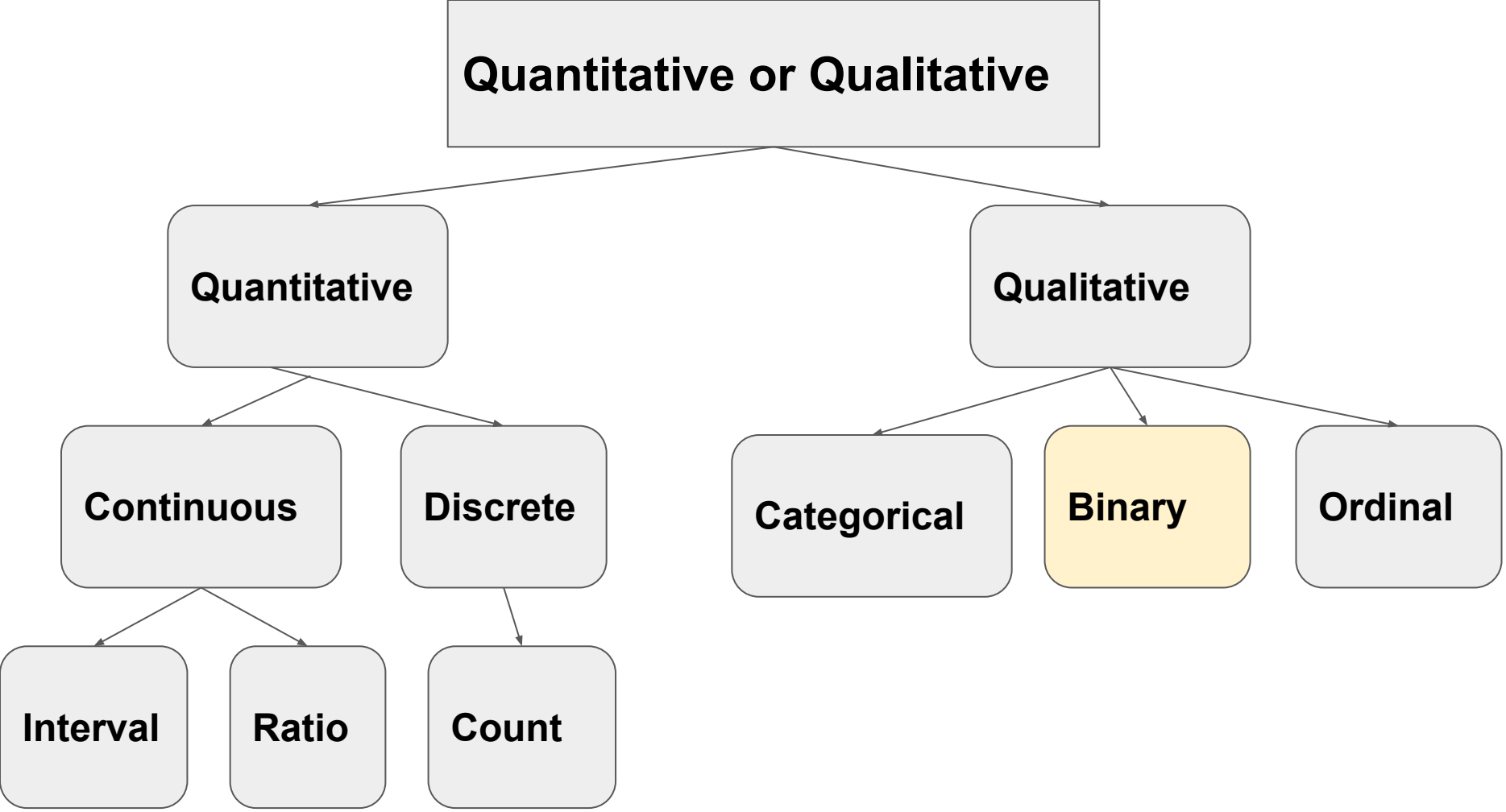

Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count

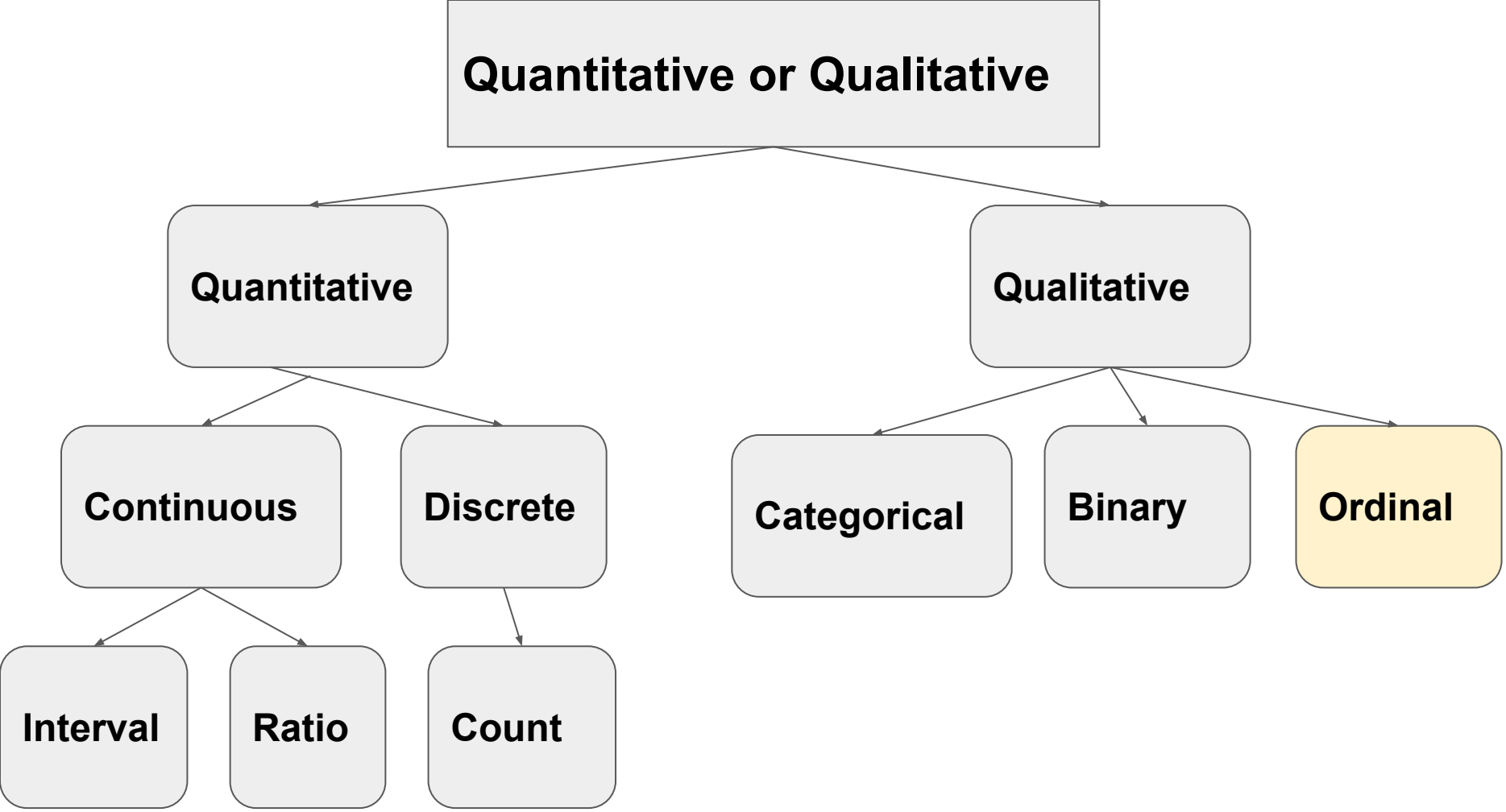

Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical / Nominal
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count


Quantitative or Qualitative
Quantitative
Qualitative
Continuous
Discrete
Categorical
Binary
Ordinal
Interval
Ratio
Count

Identify Every Column with its Data Type

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

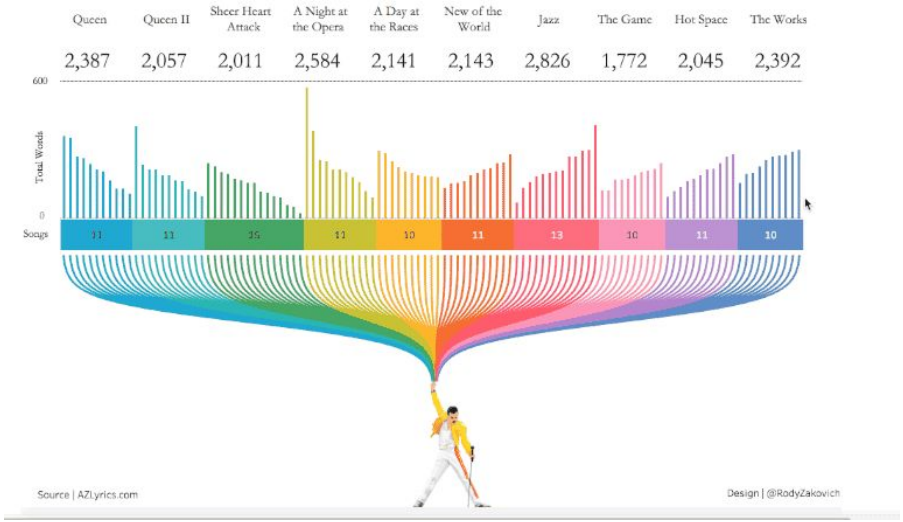
Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Topics	
data_channel_is_lifestyle: Is data channel 'Lifestyle'?	
data_channel_is_entertainment: Is data channel 'Entertainment'?	
data_channel_is_bus: Is data channel 'Business'?	
data_channel_is_socmed: Is data channel 'Social Media'?	
data_channel_is_tech: Is data channel 'Tech'?	
data_channel_is_world: Is data channel 'World'?	
Type (#)	nominal

Target	Type (#)
Number of article Mashable shares	number (1)

Stage TWO of your Analysis

How to “THINK” - Analysis

- **Univariate Analysis**
- **Bivariate Analysis**
- **Correlation**



First Things First

- **If your data is Quantitative (Numeric) then for every column of interest, get the descriptive statistics.**
  - **Analyze that and see what insights you can pull PER COLUMN (Feature)**
  - **Create a Histogram or Bar Chart or Pie Chart showing the count *y-axis) to feature (x-axis)**
- **If your data is Qualitative (Categorical) then for every column of interest, get the categories.**
  - **Get the Cardinality (Unique values in the Column (Feature)**
  - **Graph the Count / sub-category**
  - **Create a crosstab of the data**

Let’s Quantify “Popularity”

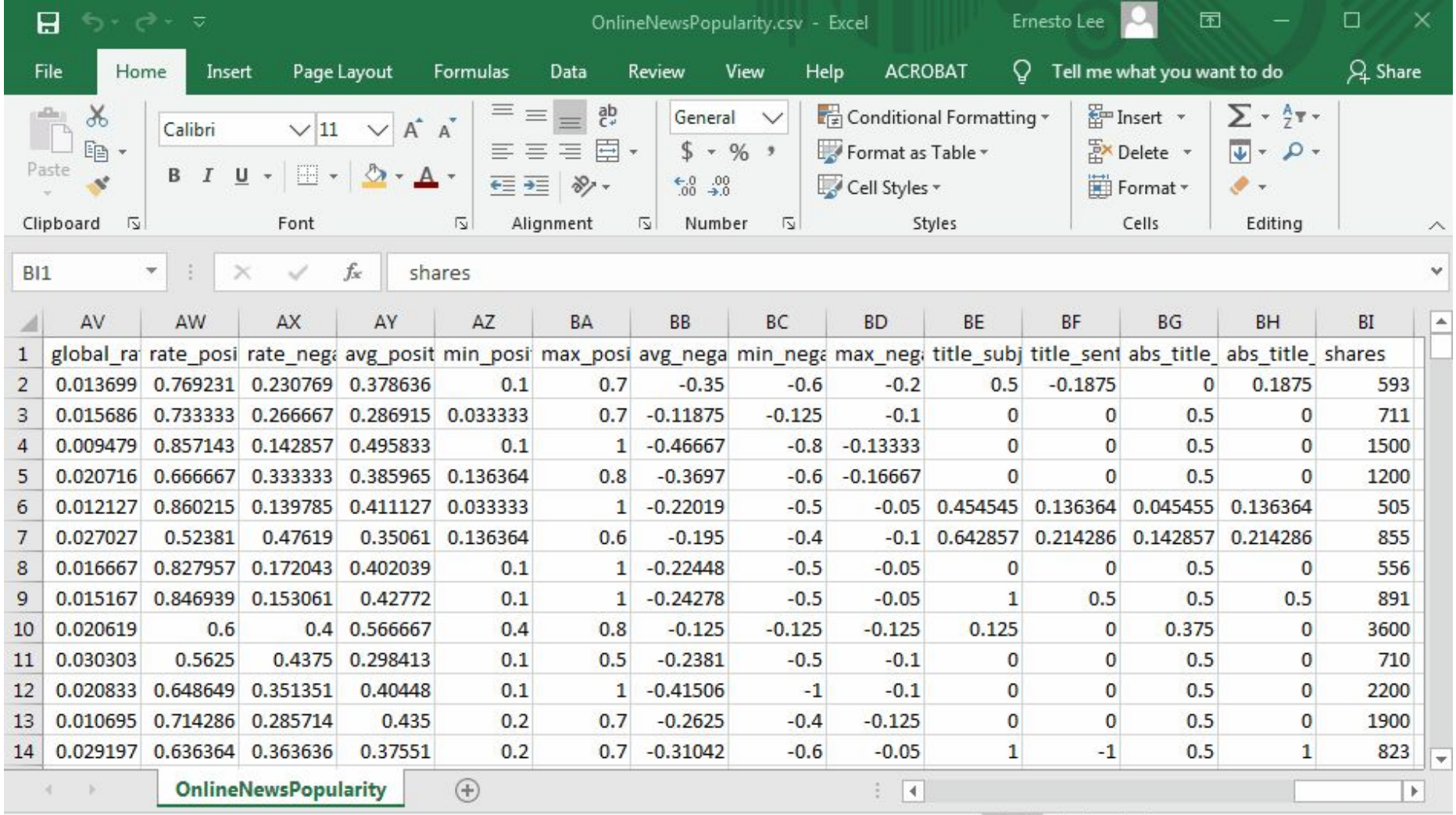
Let’s go to our SHARES feature and use statistics to define a cutoff for popularity.

We want a data driven solution so we will use either:

MEAN

MEDIAN*

MODE



	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI
1	global_ra	rate_posi	rate_nega	avg_posit	min_posi	max_posi	avg_nega	min_nega	max_nega	title_subj	title_sent	abs_title_	abs_title_	shares
2	0.013699	0.769231	0.230769	0.378636	0.1	0.7	-0.35	-0.6	-0.2	0.5	-0.1875	0	0.1875	593
3	0.015686	0.733333	0.266667	0.286915	0.033333	0.7	-0.11875	-0.125	-0.1	0	0	0.5	0	711
4	0.009479	0.857143	0.142857	0.495833	0.1	1	-0.46667	-0.8	-0.13333	0	0	0.5	0	1500
5	0.020716	0.666667	0.333333	0.385965	0.136364	0.8	-0.3697	-0.6	-0.16667	0	0	0.5	0	1200
6	0.012127	0.860215	0.139785	0.411127	0.033333	1	-0.22019	-0.5	-0.05	0.454545	0.136364	0.045455	0.136364	505
7	0.027027	0.52381	0.47619	0.35061	0.136364	0.6	-0.195	-0.4	-0.1	0.642857	0.214286	0.142857	0.214286	855
8	0.016667	0.827957	0.172043	0.402039	0.1	1	-0.22448	-0.5	-0.05	0	0	0.5	0	556
9	0.015167	0.846939	0.153061	0.42772	0.1	1	-0.24278	-0.5	-0.05	1	0.5	0.5	0.5	891
10	0.020619	0.6	0.4	0.566667	0.4	0.8	-0.125	-0.125	-0.125	0.125	0	0.375	0	3600
11	0.030303	0.5625	0.4375	0.298413	0.1	0.5	-0.2381	-0.5	-0.1	0	0	0.5	0	710
12	0.020833	0.648649	0.351351	0.40448	0.1	1	-0.41506	-1	-0.1	0	0	0.5	0	2200
13	0.010695	0.714286	0.285714	0.435	0.2	0.7	-0.2625	-0.4	-0.125	0	0	0.5	0	1900
14	0.029197	0.636364	0.363636	0.37551	0.2	0.7	-0.31042	-0.6	-0.05	1	-1	0.5	1	823

Let’s IMPUTE the Data

W	X	Y	Z	AA	AB
weekday_is_sunday	is_weekend	shares	Popularity	Serial Number	Popularity
0	0	1	Un-Popular	1	0
0	0	4	Un-Popular	2	0

 # 0 – Un-Popular
 # 1 – Popular

Now, let’s CREATE two new columns:

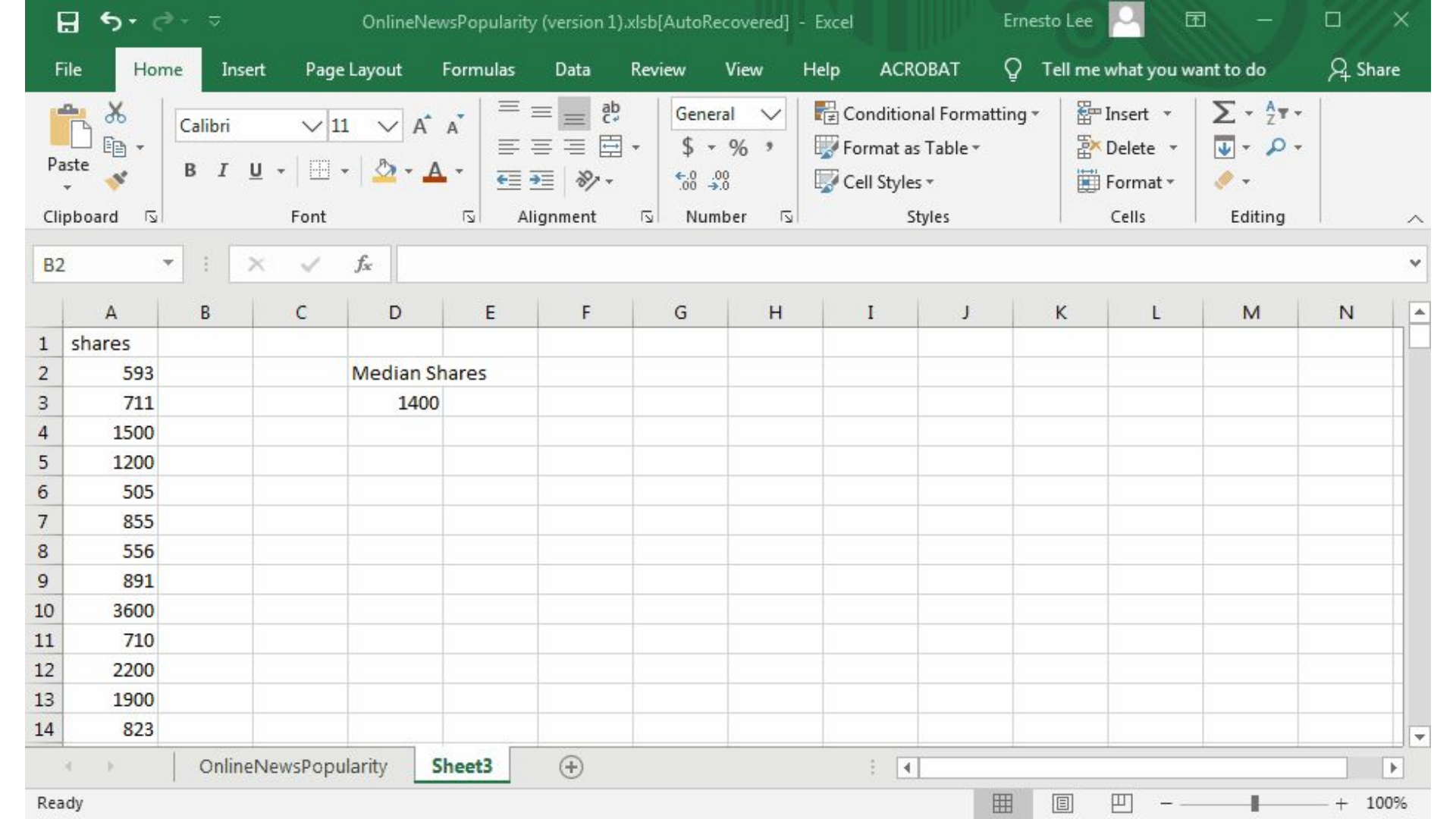
Popularity

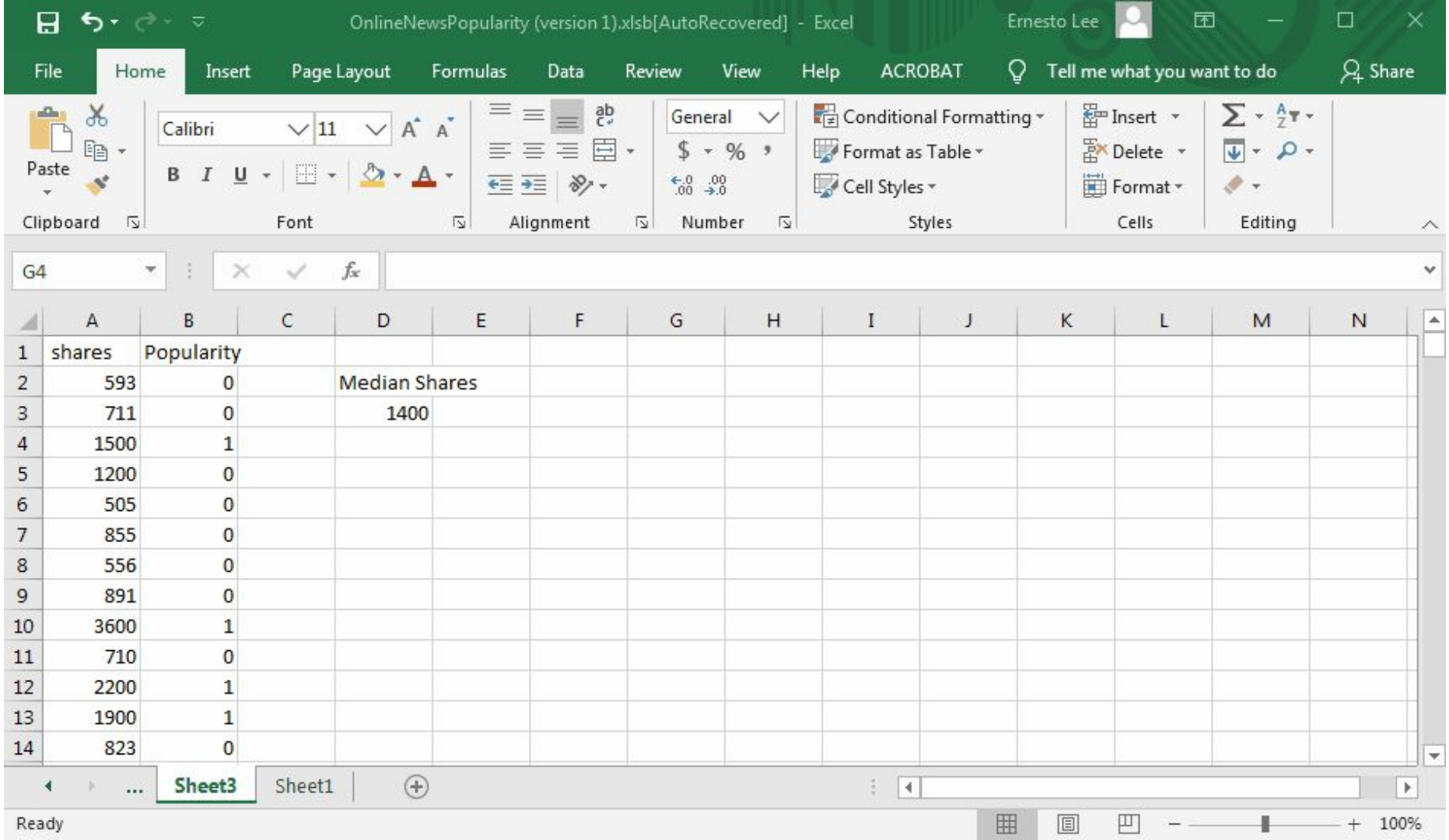
And

Encoded_Popularity

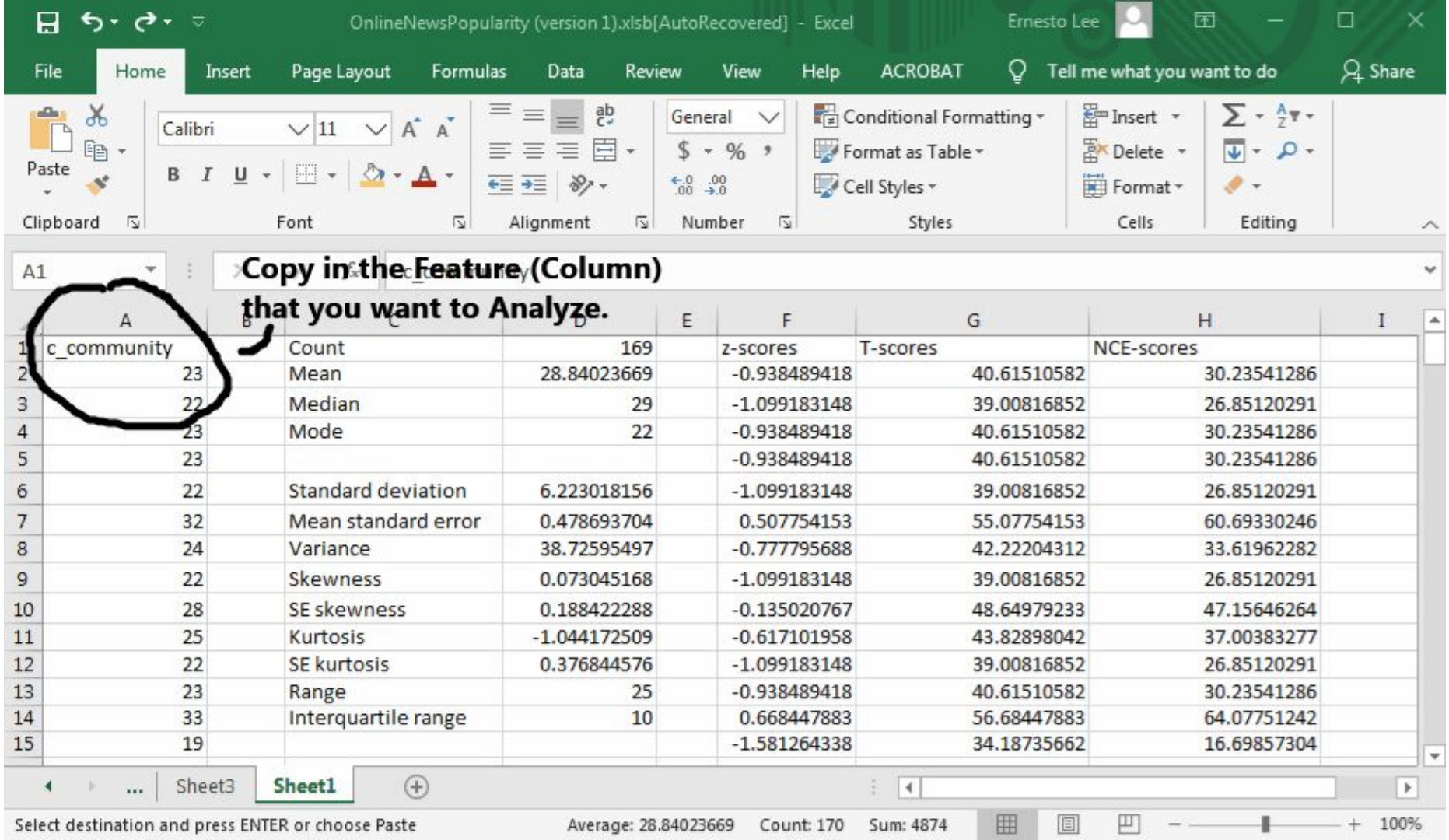
(Over 1,400 = Popular)

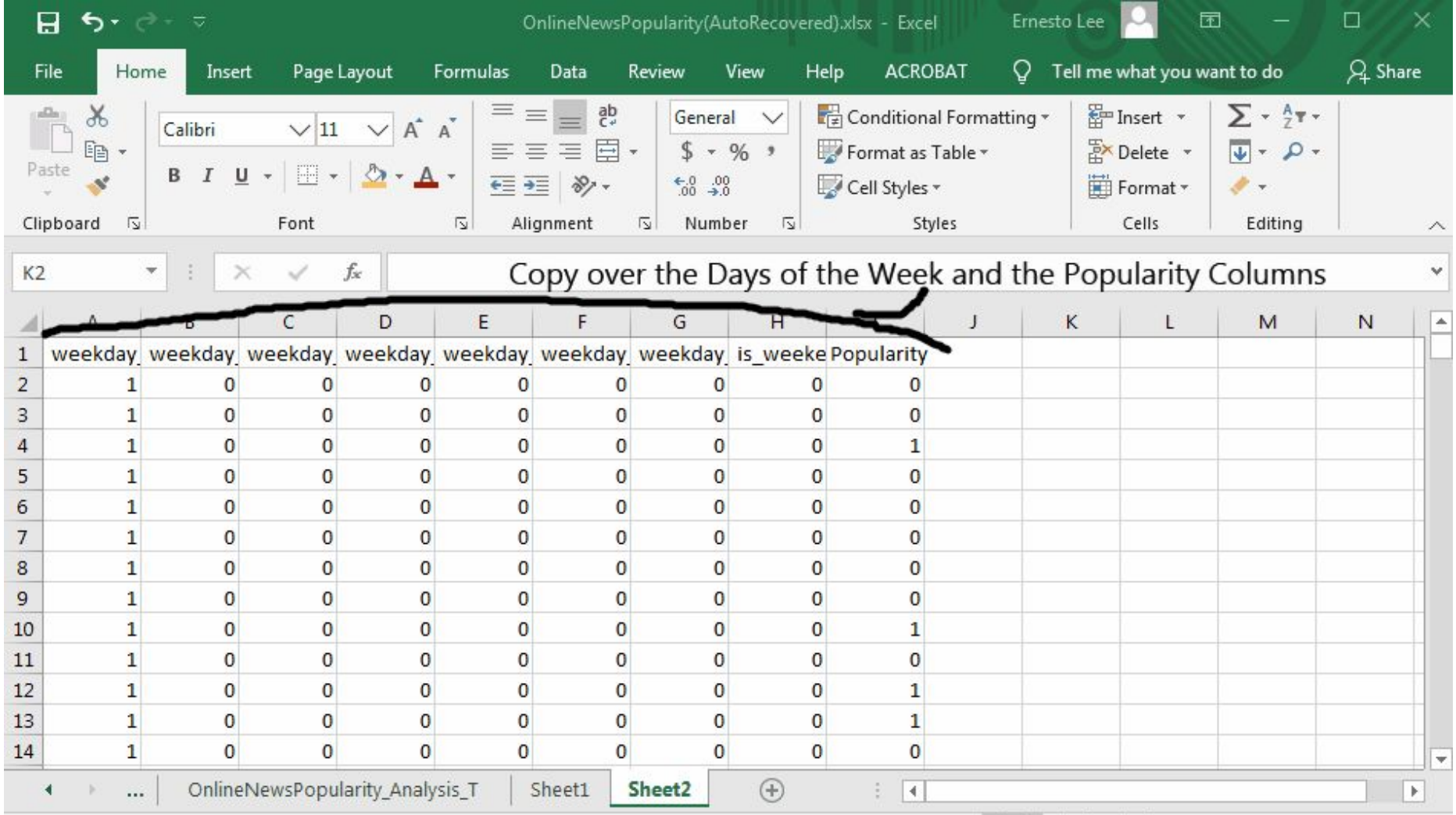
(Under 1,400=Un-popular


OnlineNewsPopularity (version 1).xlsb[AutoRecovered] - Excel
Ernesto Lee
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
ACROBAT
Tell me what you want to do
Share
Paste
Calibri
11
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
B2
shares
593
711
1500
1200
505
855
556
891
3600
710
2200
1900
823
Median Shares
1400
OnlineNewsPopularity
Sheet3
Ready
100%



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	shares	Popularity												
2	593	0		Median Shares										
3	711	0		1400										
4	1500	1												
5	1200	0												
6	505	0												
7	855	0												
8	556	0												
9	891	0												
10	3600	1												
11	710	0												
12	2200	1												
13	1900	1												
14	823	0												

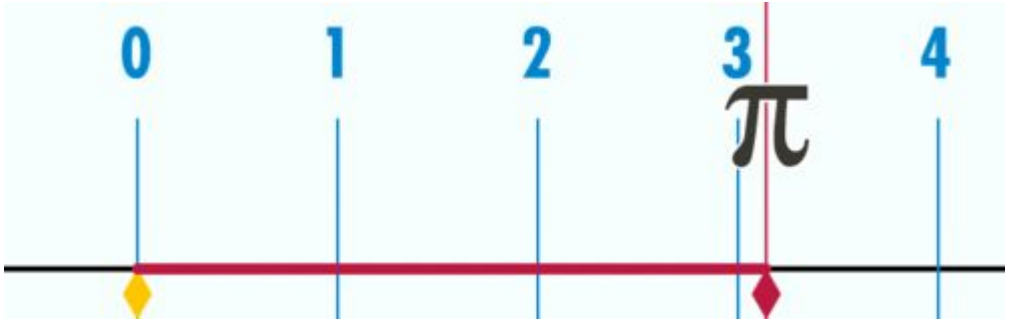

OnlineNewsPopularity (version 1).xlsb[AutoRecovered] - Excel
Ernesto Lee
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Help
ACROBAT
Tell me what you want to do
Share
Paste
Calibri
11
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
A1
Copy in the Feature (Column)
that you want to Analyze.
c_community
Count 169
Mean 28.84023669
Median 29
Mode 22
Standard deviation 6.223018156
Mean standard error 0.478693704
Variance 38.72595497
Skewness 0.073045168
SE skewness 0.188422288
Kurtosis -1.044172509
SE kurtosis 0.376844576
Range 25
Interquartile range 10
z-scores
T-scores
NCE-scores
-0.938489418 40.61510582 30.23541286
-1.099183148 39.00816852 26.85120291
-0.938489418 40.61510582 30.23541286
-0.938489418 40.61510582 30.23541286
-1.099183148 39.00816852 26.85120291
0.507754153 55.07754153 60.69330246
-0.777795688 42.22204312 33.61962282
-1.099183148 39.00816852 26.85120291
-0.135020767 48.64979233 47.15646264
-0.617101958 43.82898042 37.00383277
-1.099183148 39.00816852 26.85120291
-0.938489418 40.61510582 30.23541286
0.668447883 56.68447883 64.07751242
-1.581264338 34.18735662 16.69857304
Sheet3
Sheet1
Select destination and press ENTER or choose Paste
Average: 28.84023669
Count: 170
Sum: 4874
100%



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	weekday_	weekday_	weekday_	weekday_	weekday_	weekday_	weekday_	is_weeke	Popularity					
2	1	0	0	0	0	0	0	0	0					
3	1	0	0	0	0	0	0	0	0					
4	1	0	0	0	0	0	0	0	1					
5	1	0	0	0	0	0	0	0	0					
6	1	0	0	0	0	0	0	0	0					
7	1	0	0	0	0	0	0	0	0					
8	1	0	0	0	0	0	0	0	0					
9	1	0	0	0	0	0	0	0	0					
10	1	0	0	0	0	0	0	0	1					
11	1	0	0	0	0	0	0	0	0					
12	1	0	0	0	0	0	0	0	1					
13	1	0	0	0	0	0	0	0	1					
14	1	0	0	0	0	0	0	0	0					

Qualitative - Cardinality

If your data is Qualitative (Ordinal, Binary or Categorical), then identify the number of *unique* values in the feature (column).



OnlineNewsPopularity (version 1).xlsb[AutoRecovered] - Excel

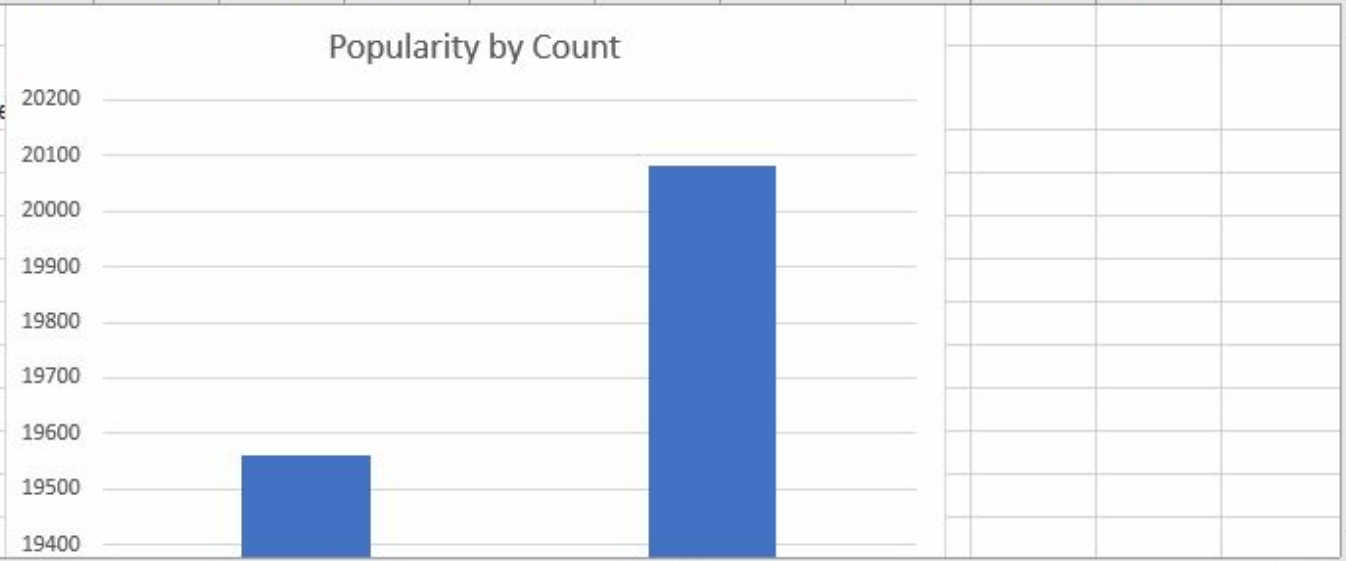
Ernesto Lee

File Home Insert Page Layout Formulas Data Review View Help ACROBAT Tell me what you want to do Share

Paste Clipboard Calibri 11 Font Alignment General Number Conditional Formatting Format as Table Cell Styles Styles Insert Delete Format Cells Editing

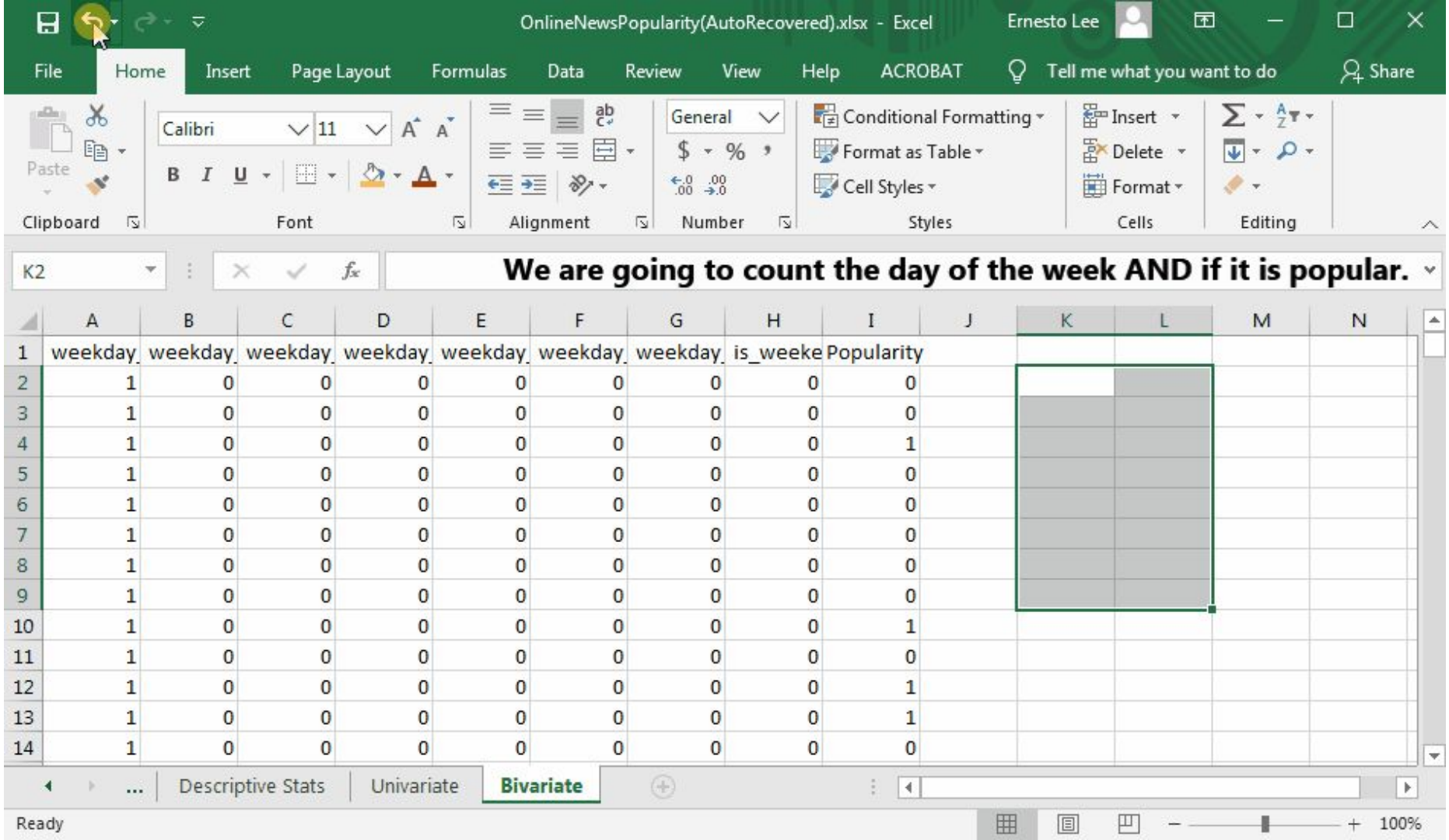
C1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	shares	Popularity												
2	593	0		Me										
3	711	0												
4	1500	1												
5	1200	0												
6	505	0												
7	855	0												
8	556	0												
9	891	0												
10	3600	1												
11	710	0												
12	2200	1												



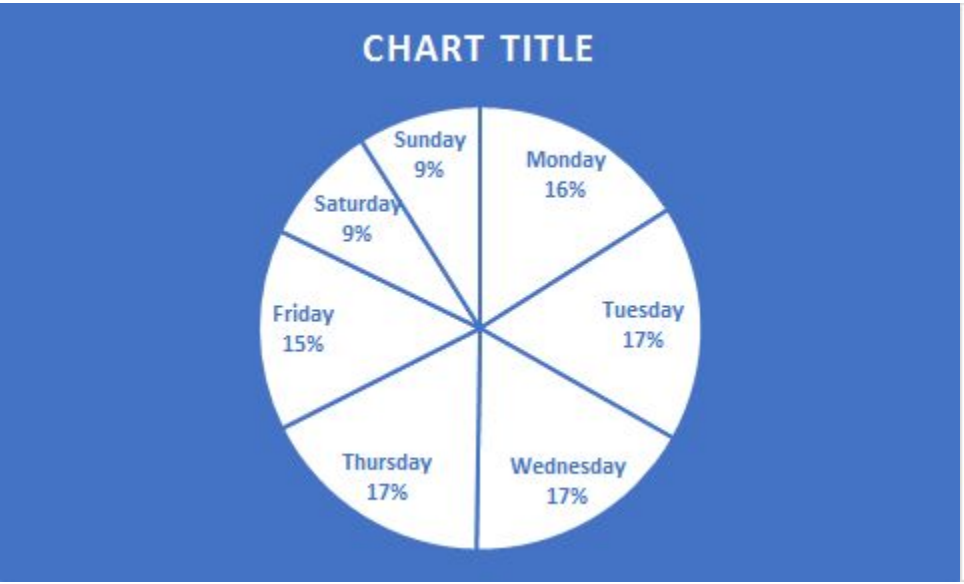
... Sheet3 Sheet1

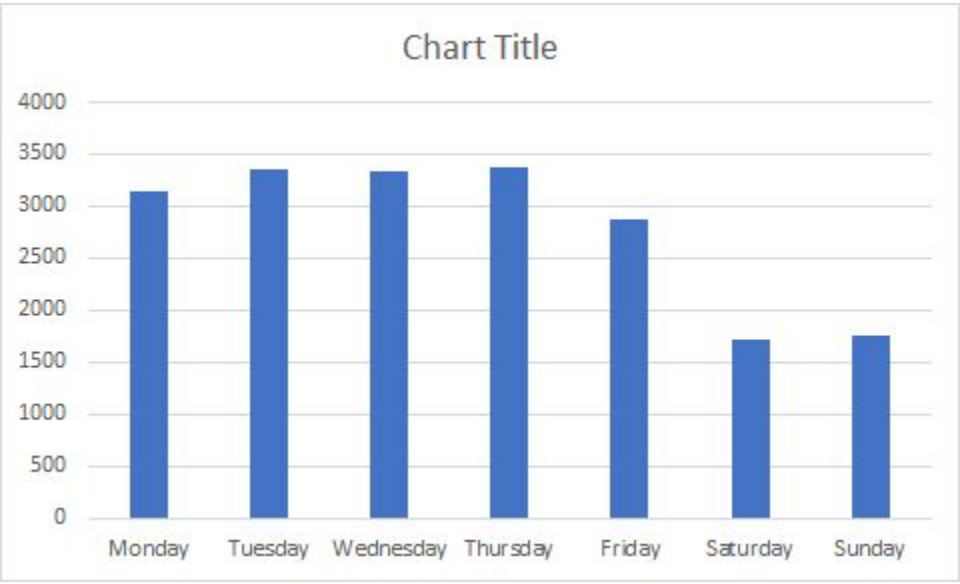
100%



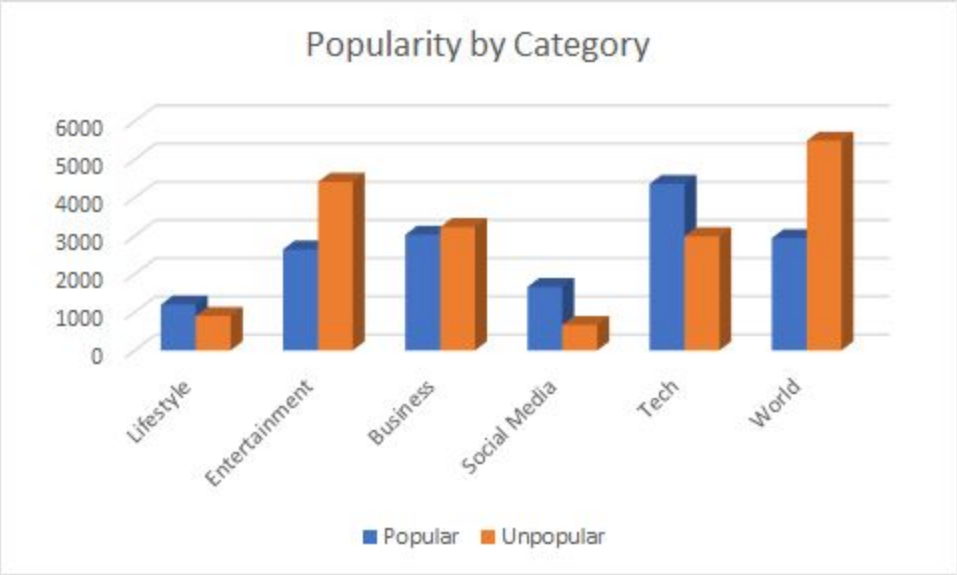
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	weekday_	weekday_	weekday_	weekday_	weekday_	weekday_	weekday_	is_weeke	Popularity					
2	1	0	0	0	0	0	0	0	0					
3	1	0	0	0	0	0	0	0	0					
4	1	0	0	0	0	0	0	0	1					
5	1	0	0	0	0	0	0	0	0					
6	1	0	0	0	0	0	0	0	0					
7	1	0	0	0	0	0	0	0	0					
8	1	0	0	0	0	0	0	0	0					
9	1	0	0	0	0	0	0	0	0					
10	1	0	0	0	0	0	0	0	1					
11	1	0	0	0	0	0	0	0	0					
12	1	0	0	0	0	0	0	0	1					
13	1	0	0	0	0	0	0	0	1					
14	1	0	0	0	0	0	0	0	0					

Article Popularity by Day of Week

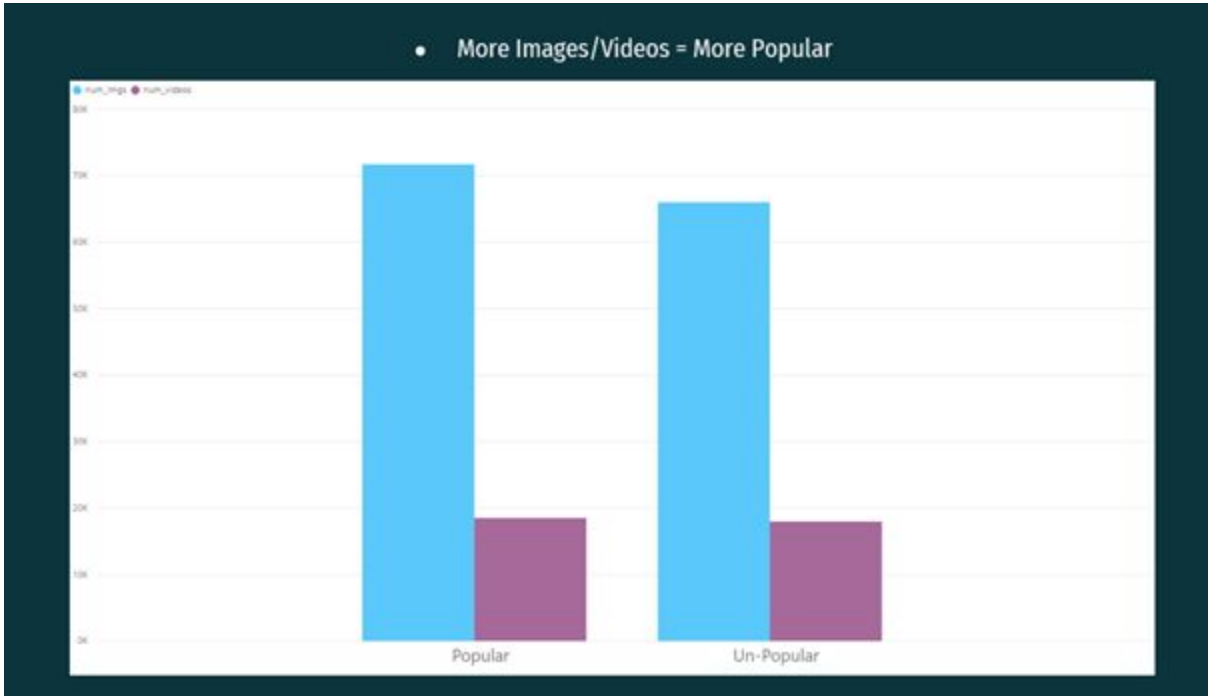




Popularity by Category



Popularity Based on Images and Videos



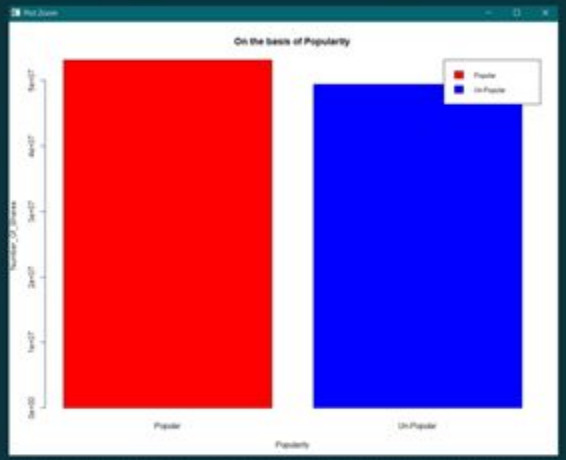
Popularity Based on Content

Content

 # >=400 – Popular
 # <400 – Un-Popular

```
> median(Scat$n_tokens_content)
[1] 400
```

- **Popularity on the bases of Content (Short and Long)**




```
C_short = (AREAA(content_short)
C_long = (AREAA(content_long)
C_short
C_long

AREAA

range=("red","blue")
Popularity=c("Popular","Un-Popular")
values= c(15076409,49931912)
barplot(values,names.arg=Popularity, xlab= "Popularity",ylab= "number_of_shares",col = range,main="On the basis of Popularity",beside=TRUE)
legend("topright",Popularity,cex=0.8, fill = range)
```

Conclusions

What have you learned?