Micro-module D: Big Data Analytic

Big data is an extremely large volume of data and datasets that come in diverse forms and from multiple sources. Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. Module D 'Big Data Analytic' is consist of 2 micro modules, which aim to provide beginner-friendly tutorials for students about data filtering and cleaning and statistical analysis.

1. Big Data and Big Data Analytic

Big Data is the term describing large sets of diverse data – structured, unstructured, and semi-structured – that are continuously generated at a high speed and in high volumes. Not only does Big Data apply to the huge volumes of continuously growing data that come in different formats, but it also refers to the range of processes, tools, and approaches used to gain insights from that data.

Big Data analytics encompasses the processes of collecting, processing, filtering/cleansing, and analyzing extensive datasets.



2. Micro-module D-D1: Data Filter and Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

In this module, you will learn the basic procedures and steps to do data cleaning in Excel, including editing data format, selecting and replacing the blank values, trimming the space, filtering data, and conditional formatting.

Introduction of Data Cleaning

What are the different types of data issues?

Duplicate data: There are 2 or more identical records. This may cause misrepresentation of inventory counts/duplication of marketing collateral or unnecessary billing activities.

Conflicting Data: When there are some records with different attributes, it means data is conflicting. For example, a company with different versions of addresses may cause delivery issues.

Incomplete Data: The data that has missing attributes. Payrolls of employees may not be processed due to their missing social security numbers in the database.

Invalid Data: Data attributes are not conforming to standardization. For example, 9-digit phone number records rather than 10 digits.

DATA CLEANING CHECKLIST



Tools for Data Cleaning

R offers a wide range of options for dealing with dirty data. The collection of packages known as the tidyverse, and adjacent packages that take a "tidy" approach, provide a range of functionality. From importing to cleaning to reshaping, these packages can help you quickly and efficiently clean messy data.

		Remove Rows
EXCE	-	One way to deal with empty cells is to remove rows that contain empty cells. This is usually OK, since data sets can be very big, and removing a few rows will not have a big impact on the result.
quick data-cleaning to	January 2020 Microsoft Excel. Below is a brief overview of five	Example Return a new Data Frame with no empty cells:
instructions to implement the solutions.	How?	<pre>import pandas as pd df = pd.read_csv('data.csv')</pre>
1 Identify all cells that contain a specific word or (short) phrase in a column with open-ended text	Use Conditional Formatting	<pre>new_df = df.dropna() print(new_df.to_string())</pre>
2 Identify and remove duplicate data	Use Remove Duplicates function or Conditional Formatting	Try it Yourself >
3 Identify the outliers within a data set (e.g., dates or grades)	Use Data Validation function	Werdy the pattern within 1 data with 1 1 are not in the rate of 1 are not interesting to 1 are not interes
Separate data from a single column into two or more columns (e.g., first and last names)	Use Flash Fill	Approx due to registration of the second secon
Categorize data in a column, such as class assignments or subject groups	Use Formula to fill in the category column	Whitevery of chitesery of the second se

While you can certainly do data cleaning in Excel, switching to R enables you to make your work reproducible. Similarly, we can use Python's Pandas and NumPy libraries to to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.

Return a new Data Fran import pandas as pd df = pd.read_csv('da			
df = pd.read_csv('da	ta.csv')		
	ta.csv')		
<pre>new_df = df.dropna()</pre>			
print(new df.to stri			
princ(new_ur.to_stri	ng())		
Try it Yourself >			
	E contrato charge an el char e Canada e		runne Heav Generates Dr.
3 Identify the outliers within a data set analisi	No	Wester Marin Sol 1 (P	Si Presentener 1 d
graden)	Hat the Proof of t	Quer: 1041 \$1.1.0	Sil (Proper Gener 1 &
grades)	PLE 25 4 Plan PLE PLE 25 Plane PLE	Quer: 1041 \$1.1.0	Sal (Proper America &
grades)	Ale Property Constraints and a second sec	Spears	Sil (Proper Gener 1 &
grades) Geparate data from a single column is more columns (e.g., first and led nam	Ale Control (Control (Contro) (Contro) (Control (Contro) (Contro) (Contro) (Contro) (Contro)	Grant Mari Si Di Si Grant Si Di Si Grant Si Di Si Si Di Si Di Si Di Si Di Si Si Di Si Di S	Sil (Proper Gener 1 &
 grades) Separate data from a single column à more columns (r.g., first and lect nam Categorier data is a solumn, such as 	Alian Control Con	Grant Mari Si Di Si Grant Si Di Si Grant Si Di Si Si Di Si Di Si Di Si Di Si Si Di Si Di S	Sil (Proper Gener 1 &
grades) Separate data from a single column is more columns (e.g., first and led nam	16	Grant Mari Si Di Si Grant Si Di Si Grant Si Di Si Si Di Si Di Si Di Si Di Si Si Di Si Di S	Sil (Proper Gener 1 &
 grades) Separate data from a single column à more columns (r.g., first and lect nam Categorier data is a solumn, such as 	Fig. 2 2 3 4 5 5 4 5 5 5 6 5 5 6 5 5 5 6 5 5 5 6 5 5 5 6 5 6 5 <td>The second secon</td> <td>Sil (Proper Gener 1 &</td>	The second secon	Sil (Proper Gener 1 &
 grades) Separate data from a single column à more columns (r.g., first and lect nam Categorier data is a solumn, such as 	Comparing the second seco	The second secon	Sil (Proper Gener 1 &
grades) grades) Separate data from a single column is more column (e.g., first and late) nam Categorize data in a column, such so assignments ar subject proups	Fig. C (L) (Z) (L) (L) (L) (L) (L) (L) (L) (L) (L) (L	Uptor	Si Proportanon : d' marconauxes : se anno : 3 ans of 5 annales se Table Tables : Table : 1000
 grades) Separate data from a single column à more columns (r.g., first and lect nam Categorier data is a solumn, such as 	Fig. C (L) (Z) (L) (L) (L) (L) (L) (L) (L) (L) (L) (L	Next to a science science science of the science of	Page face : diameter : diameter: diameter: diameter: diameter: diameter:
gudes) Sparste data tron a single column b more columno (e.g., first and last run Categorize data is a subure, such as assignments or collect groups What: Identify all cols that careta ended toxt	State	Next to a science science science of the science of	Reportance : description : descripti
guides guides guides guides guides Graphine deals have a single-optiones more originations/org. (51% and bold an an Good and a single-option and and and an Good and a single-option and and and an more official and a single-option and and and an more official and and and and and and an more official and and and and and and an more official and and and and an	Her	Went was a first and a second	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
gastre gastre	Mail The second light Mail The second light <t< td=""><td>Wenter when the second second</td><td>Reportance : description : descripti</td></t<>	Wenter when the second	Reportance : description : descripti
grotei grotei grotei fagrante data hora a single endana la mere colonaria (e.g., lost and har nam Gragmane data in a solone, such as categories data is a solone, such as defective and the second second second second What: Merelly all cells that conta defective Contributed Formersting 1. signification solution that is. In the new tab. (cited one time that (cited one data))	Mail The second se	Went was a first and a second	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
grouts grouts	44	Wenter when the second	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
grandia grandia for each of the Norse's shared deal how as shared and how the shared deal how as a suggestional star and an explosition of the shared deal how a suggestional star and each of the shared deal how a start of the shared de	Main The second se	None None None None None None None None None None None None None None None None <td>S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement</td>	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
grouts grouts		Wenter when the second	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
guidei Server de la hors rande colones a merculanto (g., for each dors random) Server de la hors random (g. for each dors random) Server de la hors random (g. for each dors random) Wart: Usen/Fy all cols has a constant and a la horse all, for each dorse all horse all h	No control of the second secon	None None None None None None None None None None None None None None None None <td>S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement</td>	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
granti grantia Grantia de la herra radio colores a Grantia de la herra radio colores a Grantia de la herra radio colores a Grantia de la herra de la herra de la herra Grantia de la herra de la herra de la herra Grantia de la herra de la herra de la herra Grantia de la herra de la herra de la herra Marcía la constituer de resenting Secon ference d'anté dans Secon ference d'anté dans Secon ference d'anté dans Any color the constituer de la desta ten- a de la marcía de la de la herra de la desta ten- de la desta ten- desta ten- de la desta ten- de la desta ten-	Note Total State Sta	Next of the second seco	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
genet genet Generation for the second and the second		None None None None None None None None None None None None None None None None <td>S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement</td>	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement
granti Secret de la forma a reflectadores Secret de la forma de l		Next of the second seco	S) Provident Same I & Sectorements - Sectorements - Sectorements - Sectorements - Same of Same Sectorements - Same Sectorement

Trim the Space

												思凡	8	- 🗉	
	Home Insert Page Layout Form				百度网島			u want to d							A Shar
L) e				o Text ge & Center	Gener Gener			aditional Fo natting ~ T Sty	able - Styles		Delete - 📑		Find & Select ~	保存到 百度网盘 保存	
5	· c? · •														
	A B C erty_I web_URL address	D	E	F Lat .7	G Lng .T	H room	l sold_price	J	K SA2	L GFA1	M GFA2	N	0	Р	
prop	5 https://hk.i8 Yan King Road	8 Yan King Road	2021/11/19				3 10,150,000		15,402/ft ²	926 ft ²	10,961/ft ²				
	6 https://hk.11 Tong Chun Street	11 Tong Chun Street	2021/11/19	22.3075	114.2624	NA	6,800,000	591 ft ²	11,506/ft ²						
	7 https://hk.<31 Razor Hill Road	31 Razor Hill Road	2021/11/19	22.33731	114.2492	NA	12,520,000	842 ft ²	14,869/ft ²	1,045 ft ²	11,981/ft²				
	8 https://hk.421 Hiu Kwong Street	21 Hiu Kwong Street	2021/11/19	22.32173	114.2289	NA	3,900,000	540 ft²	7,222/ft²	743 ft ²	5,249/ft²				
	27 https://hk.1 Po Lam Road North	1 Po Lam Road North	2021/11/18	22.31973	114.253	NA	5,180,000	588 ft²	8,810/ft ²	652 ft ²	7,945/ft²				
	29 https://hk.i9 Tong Tak Street	9 Tong Tak Street	2021/11/18	22.3072	114.2571		2 8,500,000	465 ft²	18,280/ft ²	617 ft ²	13,776/ft ²				
	31 https://hk.i21 Hiu Kwong Stree	t 21 Hiu Kwong Street	2021/11/18	22.32138	114.2283	NA	3,980,000	433 ft ²	9,192/ft²	596 ft ²	6,678/ft²				
	32 https://hk.1 Ngan O Road	1 Ngan O Road	2021/11/18	22.31386	114.2658		2 7,900,000	436 ft²	18,119/ft²	606 ft ²	13,036/ft ²				
	33 https://hk.i13 Laguna Street	13 Laguna Street	2021/11/18	22.30491	114.2268		2 8,200,000	517 ft²	15,861/ft²	644 ft ²	12,733/ft ²				
	34 https://hk.i13 Laguna Street	13 Laguna Street	*2021/11/18	22.30491	114.2268		3 11,500,000	748 ft²	15,374/ft ²	939 ft ²	12,247/ft ²				
	35 https://hk.11 Lei Yue Mun Road	11 Lei Yue Mun Road	2021/11/18	22.30491	114.2268		2 5,200,000	506 ft ²	10,277/ft ²	677 ft ²	7,681/ft ²				
	36 https://hki20 Shan Kwong Road		2021/11/18	22.30491	114.2268		3 36,000,000	1,531 ft²	23,514/ft ²	1,955 ft ²	18,414/ft ²				
	Sheet1 (+)									11C					

3. Micro-module D-D2: Statistical Data Analysis

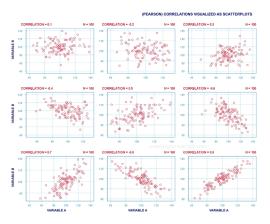
SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others.

We usually use SPSS to input, manage, analysis a large amount of quantitative data in the urban studies field. This tutorial will show the basic concepts for beginners to manage their data in SPSS, including the introduction of SPSS interface, data input, and categories of variables, it will also include the descriptive statistics step-by-step guidelines.

Introduction of SPSS

What is SPSS?

The program, originally called Statistical Package for the Social Sciences, was released in 1968 and quickly became one of the most widely used statistics programs in the social sciences, including in healthcare, government, market research and surveying.



SPSS is GREAT for

1) Opening data files, either in SPSS' own file format or many others;

 editing data such as computing sums and means over columns or rows of data. SPSS has outstanding options for more complex operations as well.

 wore complex operations as well.
 creating tables and charts containing frequency counts or summary statistics over (groups of) cases and variables.
 running inferential statistics such as ANOVA, regression

and factor analysis.saving data and output in a wide variety of file formats.

https://www.spss-tutorials.com/spss-correlation-analysis/

Variable Type

NUMERIC

Numeric variables, as you might expect, have data values that are recognized as numbers. This means that they can be sorted numerically or entered into arithmetic calculations. When viewed in the Data View window, system-missing values for numeric variables will appear as a dot (i.e.,

STRING

Use when you want to type letters. For example, peoples' names, breeds of dog, occupations. You can also include numbers or symbols, but they will be treated by SPSS as text. For example, zip codes are numeric but you may want to treat them as text.

СОММА

Numeric variables that are separated every three places by a comma. For example, 100,000.00 or 999,988,565.21.

DOT

Similar to comma, but the dot is used to separate the three places and a comma is used to indicate a decimal. For example. 100.000,00 and 999.988.565,21. Not used in the UK or USA, but common in some other countries.

Frequencies Option

			·		*											
															Visible: 5 of	f 5 Varia
	astreetsection	/ pedestria /	yactivites	pionalac d tivities	e sds var	var	var	var	var	var	var	var	var	var	var	va
1	Tai Po Road #1	11	5	6	2											
2	Tai Po Road #2	4	4	0	2											
3	Tai Po Road #3	4	4	0	3											
4	Tai Po Road #4	2	the Francisco	cies: Statistics				×								
5	Sha Tin Centre St #1	34	Trequer					~								
	Sha Tin Centre St #2	16	Percentil			Central Tende										
	Sha Tin Centre St #3	10	Quart	les		🖌 Mean										
	Sha Tin Centre St #4	12	Cut p	oints for: 10	equal groups	Megian										
	Sha Tin Centre St #5	21	Perce	ntile(s):		Mode										
	Sha Tin Centre St #6	27	A			🖉 gum										
11	Sha Tin Centre St #7	7	Cha													
	Tam Kon Pon St #1	15														
	Tam Kon Pon St #2	25	Rem	love.												
	Yuen Wo Road #1	5														
	Yuen Wo Road #2	20				🛛 Vajues are	group midpois	its								
	Warn Pok St #1	21	Dispersie	in		Distribution -										
17	Wam Pok St #2	27	E Std. d	eviation 📃 Mij	nimum	Skewness										
	Warn Pok St #3	24	🖾 Variar	ice 🛅 Ma	ajmum	🖾 Kurtosis										
	Pak Hok Tin St #1	8	E Rang	e 🖪 S.	E. mean											
	Pak Hok Tin St #2	4		-												
	Pak Hok Tin St #3	3		20	ntinue Cancel	Help	Help									
	Tin Ho Road #1	9	a	U	1											
	Tin Ho Road #2	14	13	1	2											
	Tin Ho Road #3	13	13	0	1											
	Tin Mei St #1	9	6	3	2											
	Tin Mei St #2	12	12	0	2											
	Tin Yin Road #1	28	16	12	11											
28	Tin Yin Road #2	23	18	5	7											