



CPE 352 Data Science

2 – Tabular Data and EDA

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi



Topics

- What is tabular data?
- Forms of tabular data
- Exploring tabular data

<https://drive.google.com/file/d/11SD6D0sLG5GDY7cDwk0gZVFKqQSC-VyJ/view?usp=sharing>

What is tabular data?

- Tabular data, or table, is a set of data elements that represents **attributes/variables** in vertical **columns** and **units** in horizontal rows

attributes/fields/variables

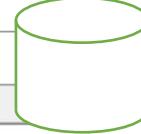
unit



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	City	State	Country	Postal Code	Market
2	32298	CA-2012-124	31/07/2012	31/07/2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	United States	10024	US
3	26341	IN-2013-778	05/02/2013	07/02/2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales	Australia		APAC
4	25330	IN-2013-7124	17/10/2013	18/10/2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	Australia		APAC
5	13524	ES-2013-1575	28/01/2013	30/01/2013	First Class	KM-16375	Katherine McHome	Office	Berlin	Berlin	Germany		EU
6	47221	SG-2013-4321	05/11/2013	06/11/2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar	Senegal		Africa
7	22732	IN-2013-4236	28/06/2013	01/07/2013	Second Class	JM-15655	Jim Mitchum	Corporate	Sydney	New South Wales	Australia		APAC
8	30570	IN-2011-8182	07/11/2011	09/11/2011	First Class	TS-21340	Toby Swinde	Consumer	Porirua	Wellington	New Zealand		APAC
9	31192	IN-2012-8636	14/04/2012	18/04/2012	Standard Class	MB-18085	Mick Brown	Consumer	Hamilton	Waikato	New Zealand		APAC
10	40155	CA-2014-135	14/10/2014	21/10/2014	Standard Class	JW-15220	Jane Waco	Corporate	Sacramento	California	United States	95823	US
11	40936	CA-2012-116	28/01/2012	31/01/2012	Second Class	JH-15985	Joseph Holt	Consumer	Concord	North Carolina	United States	28027	US
12	34577	CA-2011-102	05/04/2011	09/04/2011	Second Class	GM-14695	Greg Maxwell	Corporate	Alexandria	Virginia	United States	22304	US
13	28879	ID-2012-284C	19/04/2012	22/04/2012	First Class	AJ-10780	Anthony Jacc	Corporate	Kabul	Kabul	Afghanistan		APAC
14	45794	SA-2011-1831	27/12/2011	29/12/2011	Second Class	MM-7260	Magdelene K	Consumer	Jizan	Jizan	Saudi Arabia		EMEA
15	4132	MX-2012-13C	13/11/2012	13/11/2012	Same Day	VF-21715	Vicky Freyma	Home Office	Toledo	Parana	Brazil		LATAM
16	27704	IN-2013-7395	06/06/2013	08/06/2013	Second Class	PF-19120	Peter Fuller	Consumer	Mudanjiang	Heilongjiang	China		APAC
17	13779	ES-2014-5095	31/07/2014	03/08/2014	Second Class	BP-11185	Ben Peterma	Corporate	Paris	Ille-de-France	France		EU
18	36178	CA-2014-143	03/11/2014	06/11/2014	Second Class	TB-21175	Thomas Bola	Corporate	Henderson	Kentucky	United States	42420	US
19	12069	ES-2014-165	08/09/2014	14/09/2014	Standard Class	PJ-18835	Patrick Jones	Corporate	Prato	Tuscany	Italy		EU
20	22096	IN-2014-1176	31/01/2014	01/02/2014	First Class	JS-15685	Jim Sink	Corporate	Townsville	Queensland	Australia		APAC
21	49463	TZ-2014-819C	05/12/2014	07/12/2014	Second Class	RH-9555	Ritsa Hightov	Consumer	Uvinza	Kigoma	Tanzania		Africa
22	46630	PL-2012-782C	08/08/2012	10/08/2012	First Class	AB-600	Ann Bluma	Corporate	Bytom	Silesia	Poland		EMEA
23	31784	CA-2011-154	29/10/2011	31/10/2011	First Class	SA-20830	Sue Ann Ree	Consumer	Chicago	Illinois	United States	60610	US
24	21586	IN-2011-448C	02/05/2011	03/05/2011	First Class	JK-15325	Jason Klamcz	Corporate	Suzhou	Anhui	China		APAC
25	13528	ES-2013-2864	27/02/2013	01/03/2013	Second Class	LB-16795	Laurel Beltra	Home Office	Edinburgh	Scotland	United Kingdom		EU
26	1570	US-2014-133	31/07/2014	01/08/2014	First Class	NP-18325	Naresj Patel	Consumer	Juárez	Chihuahua	Mexico		LATAM
27	3484	MX-2014-165	05/09/2014	08/09/2014	First Class	VD-21670	Valérie Domí	Consumer	Soyapango	San Salvador	El Salvador		LATAM
28	30191	IN-2011-102E	17/12/2011	20/12/2011	First Class	PB-19210	Phillip Breyer	Corporate	Taipei	Taipei City	Taiwan		APAC

Format

- Database



Order_ID	Order_Date	Ship_Date	Ship_Mode	Customer_ID	Product_ID	Sales	Quantity	Discount	Profit
CA-2013-152156	2014-11-09	2014-11-12	Second Class	CG-12520	FUR-BO-10001798	261.96	2	0.00	41.9136
CA-2013-152156	2014-11-09	2014-11-12	Second Class	CG-12520	FUR-CH-10000454	731.9399999999999	3	0.00	219.5819999999997
CA-2013-138688	2014-06-13	2014-06-17	Second Class	DV-13045	OFF-LA-10000240	14.62	2	0.00	6.871399999999995
US-2012-108966	2013-10-11	2013-10-18	Standard Class	SO-20335	FUR-TA-10000577	957.5775	5	0.45	-383.03100000000006
US-2012-108966	2013-10-11	2013-10-18	Standard Class	SO-20335	OFF-ST-10000760	22.368000000000002	2	0.20	2.516399999999999
CA-2011-115812	2012-06-09	2012-06-14	Standard Class	BH-11710	FUR-FU-10001487	48.86	7	0.00	14.169399999999996

- Flat files (csv, tsv, txt, ...)

```
D:\Dropbox\_Teaching\Data Science\2021\Lecture 2>more 2-1-sales_transactions.csv
Row ID|Order ID|Order Date|Ship Date|Ship Mode|Customer ID|Customer Name|Segment|Country|City|State|Postal Code|Region|Product ID|Category|Sub-Category|Product Name|Sales|Quantity|Discount|Profit
1|CA-2013-152156|2014-11-09T00:00:00Z|2014-11-12T00:00:00Z|Second Class|CG-12520|Claire Gute|Consumer|United States|Henderson|Kentucky|42420|South|FUR-BO-10001798|Furniture|Bookcases|Bush Somerset Collection Bookcase|261.96|2|0|41.9136
2|CA-2013-152156|2014-11-09T00:00:00Z|2014-11-12T00:00:00Z|Second Class|CG-12520|Claire Gute|Consumer|United States|Henderson|Kentucky|42420|South|FUR-CH-10000454|Furniture|Chairs|Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back|731.9399999999999|3|0|219.5819999999997
3|CA-2013-138688|2014-06-13T00:00:00Z|2014-06-17T00:00:00Z|Second Class|DV-13045|Darrin Van Huff|Corporate|United States|Los Angeles|California|90036|West|OFF-LA-10000240|Office Supplies|Labels|Self-Adhesive Address Labels for Typewriters by Universal|14.62|2|0|6.871399999999995
4|US-2012-108966|2013-10-11T00:00:00Z|2013-10-18T00:00:00Z|Standard Class|SO-20335|Sean O'Donnell|Consumer|United States|Fort Lauderdale|Florida|33311|South|FUR-TA-10000577|Furniture|Tables|Bretford CR4500 Series Slim Rectangular Table|957.5775|5|0.45|-383.0310000000006
```

- API (may need transformation)

```
1  {
2      "success": true,
3      "timestamp": 1620386223,
4      "source": "GBP",
5      "quotes": {
6          "GBPAED": 5.109727,
7          "GBPAFN": 107.750768,
8          "GBPALL": 141.659078,
```

Forms of tabular data

Transactional table

- Each row represents the transaction
- Columns contain transaction attributes / sometime may contains larger context, e.g. customers, products
- May not be at a suitable unit of analysis
- Examples
 - Sales transaction
 - Activity transaction
 - Call center log
 - Game play log
 - Patient visit

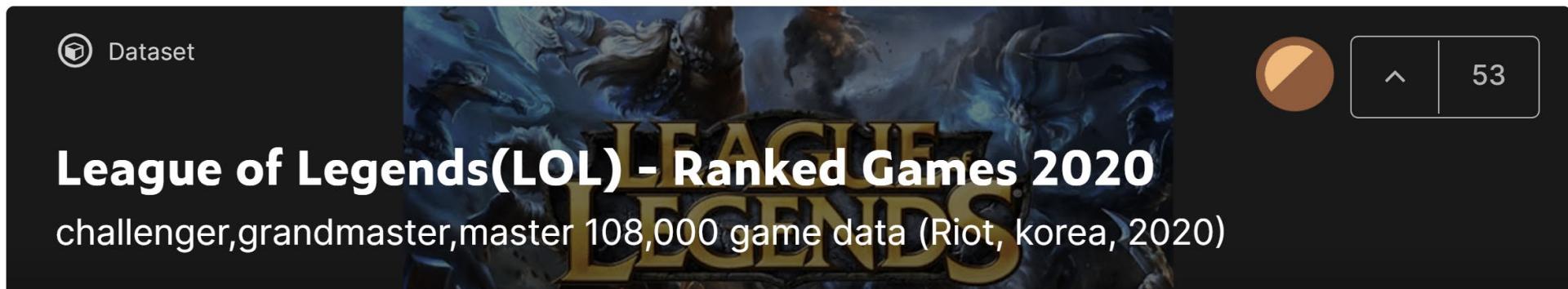


Example: sales transactions

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC						
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit													
2	1	CA-2013-152156	2014-11-09T00:00:00Z	2014-11-12T00:00:00Z	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96	2	0	41.9136													
3	2	CA-201	Rounded Back	731.939999999999	3	0	219.581999999999																											
4	3	CA-2013-138688	2014-06-13T00:00:00Z	2014-06-17T00:00:00Z	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62	2	0	6.871399999999995													
5	4	US-2012-108966	2013-10-11T00:00:00Z	2013-10-18T00:00:00Z	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-1000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.03100000000006													
6	5	US-2012-108966	2013-10-11T00:00:00Z	2013-10-18T00:00:00Z	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.36800000000002	2	0.2	2.516399999999999													
7	6	CA-201	Cherry Wood	48.86	7	0	14.16939999999996																											
8	7	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell 322	7.28	14	0	1.9656000000000002													
9	8	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-1000275	Technology	Phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.71520000000004													
10	9	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Angle-View Binders with Locking Rings by Samsill	18.504	3	0.2	5.7825													
11	10	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.46999999999999													
12	11	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10001539	Furniture	Tables	Chromcraft Rectangular Conference Tables	1706.184000000002	9	0.2	85.3091999999998													
13	12	CA-2011-115812	2012-06-09T00:00:00Z	2012-06-14T00:00:00Z	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002033	Technology	Phones	Konftel 250 ConferenceÂ phoneA - Charcoal black	911.424	4	0.2	68.35680000000002													
14	13	CA-2014-114412	2015-04-16T00:00:00Z	2015-04-21T00:00:00Z	Standard Class	AA-10480	Andrew Allen	Consumer	United States	Concord	North Carolina	28027	South	OFF-PA-10002365	Office Supplies	Paper	Xerox 1967	15.55200000000003	3	0.2	5.4432													
15	14	CA-2013-161389	2014-12-06T00:00:00Z	2014-12-11T00:00:00Z	Standard Class	IM-15070	Irene Maddox	Consumer	United States	Seattle	Washington	98103	West	OFF-BI-10003656	Office Supplies	Binders	Fellowes PB200 Plastic Comb Binding Machine	407.9760000000006	3	0.2	132.5921999999999													
16	15	US-201	Very Large HEPA Filter	68.80999999999995	[5]	0.8	-123.858																											
17	16	US-2012-118983	2013-11-22T00:00:00Z	2013-11-26T00:00:00Z	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texas	76106	Central	OFF-BI-10000756	Office Supplies	Binders	Storex DuraTech Recycled Plastic Frosted Binders	2.5439999999999996	3	0.8	-3.816000000000													
18	17	CA-201	Vertical L-Shelf	72 ¹¹ H x 36 ¹¹ W x 18 1 ¹¹ D ¹¹	[6]	65.88	[6]	[0]	13.317599999999999																									
19	18	CA-2011-167164	2012-05-13T00:00:00Z	2012-05-15T00:00:00Z	Second Class	AG-10270	Alejandro Grove	Consumer	United States	West Jordan	Utah	84084	West	OFF-ST-10000107	Office Supplies	Storage	Fellowes Super Stor/Drawer	55.5	2	0	9.989999999999995													
20	19	CA-2011-143336	2012-08-27T00:00:00Z	2012-09-01T00:00:00Z	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West	OFF-AR-10000356	Office Supplies	Art	Newell 341	8.56	2	0	2.4823999999999993													
21	20	CA-2011-143336	2012-08-27T00:00:00Z	2012-09-01T00:00:00Z	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	California	94109	West	TEC-PH-10001949	Technology	Phones	Cisco SPA 501G IP Phone	213.48000000000002	3	0.2	16.010999999999998													
22	21	CA-201	White	1 ¹¹ "x22 ¹¹ z	[4]	0.2	1.383999999999999																											
23	22	CA-2013-137330	2014-12-10T00:00:00Z	2014-12-14T00:00:00Z	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebraska	68025	Central	OFF-AR-10000246	Office Supplies	Art	Newell 318	19.459999999999997	7	0	5.0596													
24	23	CA-201	4' Cord Length	60.33999999999999	[7]	0	15.688400000000001																											
25	24	US-201	Gray	71.37199999999999	[2]	0.3	-1.019600000000005																											
26	25	CA-2012-106320	2013-09-25T00:00:00Z	2013-09-30T00:00:00Z	Standard Class	EB-13870	Emily Burns	Consumer	United States	Orem	Utah	84057	West	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	1044.629999999999	3	0	240.2649													
27	26	CA-2013-121755	2014-01-16T00:00:00Z	2014-01-20T00:00:00Z	Second Class	EH-13945	Eric Hoffmann	Consumer	United States	Los Angeles	California	90049	West	OFF-BI-10001634	Office Supplies	Binders	Wilson Jones Active Use Binders	11.648000000000001	2	0.2	4.2224													
28	27	CA-2013-121755	2014-01-16T00:00:00Z	2014-01-20T00:00:00Z	Second Class	EH-13945	Eric Hoffmann	Consumer	United States	Los Angeles	California	90049	West	TEC-AC-10003027	Technology	Accessories	ImationÂ 8GB Mini TravelDrive USB 2.0A Flash Drive	90.57000000000001	3	0	11.7741000000													
29	28	US-201	Royal Cherry Finish	3083.430000000003	[7]	0.5	-1665.052																											
30	29	US-2012-150630	2013-09-17T00:00:00Z	2013-09-21T00:00:00Z	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-BI-10000474	Office Supplies	Binders	Avery Recycled Flexi-View Covers for Binding Systems	9.61800000000002	2	0.7	-7.0532													
31	30	US-2012-150630	2013-09-17T00:00:00Z	2013-09-21T00:00:00Z	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Pennsylvania	19140	East	FUR-FU-10004848	Furniture	Furnishings	"Howard Miller 13-3/4" Diameter Brushed Chrome Round Wall Clock"	124.2000000000000	1	0.2	0.0													
32	31	US-2012-150630	2013-09-17T00:00:00Z	2013-09-21T00:00:00Z	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-EN-10001509	Office Supplies	Envelopes	Poly String Tie Envelopes	3.2640000000000002	2	0.2	1.1015999999999997													
33	32	US-201	Putty/Woodgrain	186.304	[6]	0.2	19.709199999999989																											
34	33	US-201	14 7/4" x Executive Red	16.858000000000001	[6]	1	0.7	-5.715																										
35	34	US-2012-150630	2013-09-17T00:00:00Z	2013-09-21T00:00:00Z	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Pennsylvania	19140	East	OFF-AR-10001683	Office Supplies	Art	Lumber Crayons	15.76	2	0.2	3.5460000000000007													
36	35	CA-2014-107727	2015-10-20T00:00:00Z	2015-10-24T00:00:00Z	Second Class	MA-17560	Matt Abelman	Home Office	United States	Houston	Texas	77095	Central	OFF-PA-10000249	Office Supplies	Paper	Easy-staple paper	29.472	3	0.2	9.946799999999998													
37	36	CA-2013-117590	2014-12-09T00:00:00Z	2014-12-11T00:00:00Z	First Class	GH-14485	Gene Hale	Corporate	United States	Richardson	Texas	75080	Central	TEC-PH-10004977	Technology	Phones	GE 30524EE4	1097.5440000000003	7	0.2	123.4736999999999													
38	37	CA-201	Black	190.92	[5]	0.6	-147.9630000000002																											
39	38	CA-2012-117415	2013-12-27T00:00:00Z	2013-12-31T00:00:00Z	Standard Class	SN-20710	Steve Nguyen	Home Office	United States	Houston	Texas	77041	Central	OFF-EN-10002986	Office Supplies	Envelopes	"#10-4 1/8" x 9 1/2" Premium Diagonal Seam Envelopes"	113.328	9	0.2	35.415													
40	39	CA-201	Custom Colors	532.3992	[3]	0.32	-46.9764000000001																											

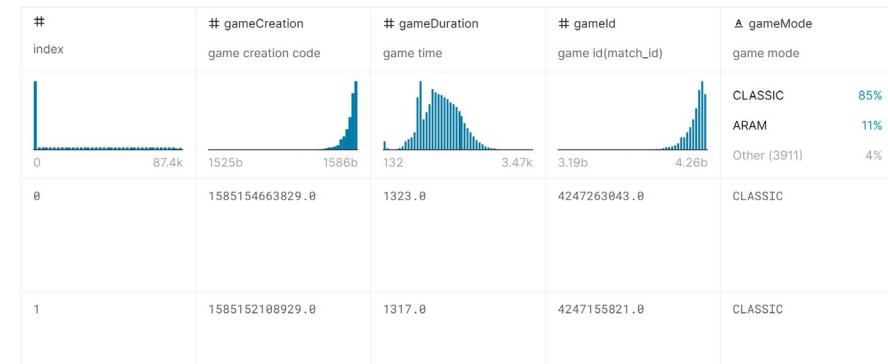
http://fastdata.in.th/HDS_DS/data/2-1-sales_transactions.csv

Example: game transaction



Introduction

- challenger, grandmaster, master 100,800 game data(korea, 2020)
- There are three large categories
 - team object data
 - participants data
 - gameDuration



<https://www.kaggle.com/gyejr95/league-of-legendslol-ranked-games-2020-ver1>

Example: COVID-19 case in Thailand, case report

ข้อมูลข่าวสาร กลุ่มข้อมูล

รายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย

องค์กร : กรมควบคุมโรค

 ดาวน์โหลดกับหมด

DGA

DGA หรือ สำนักงานพัฒนารัฐบาลดิจิทัล (องค์กรมหาชน) ได้เปิดตัวบริการดูแลข้อมูล “รายงาน COVID-19 ประจำวัน ข้อมูลประจำวันประเทศไทย” บนเว็บไซต์ DATA GO TH ซึ่งให้บริการดาวน์โหลดข้อมูลในรูปแบบไฟล์ XLSX และ CSV รวมทั้งต้องการเข้าถึงข้อมูลเพื่อใช้กับ API ของ DGA ที่ต้องการใช้บริการต่อไปโดยไม่ต้องเขียนโค้ดเอง เพื่อเข้ามายังระบบของตน ทั้งนี้ต้องขอสงวนสิทธิ์ไม่สามารถนำข้อมูลนี้ไปเผยแพร่ในสื่อโซเชียลมีเดียได้

DGA ขอเชิญชวนผู้อ่านที่สนใจสามารถเข้ามาใช้งานเว็บไซต์ DATA GO TH และตรวจสอบข้อมูลที่ต้องการได้ตามลิงค์ด้านล่าง

<https://data.go.th/dataset/covid-19-daily>

รายงานผู้ป่วยรายบันประจำวัน จาก กรมควบคุมโรค

สามารถเดินทางสถานการณ์ได้ที่ <https://ddc.moph.go.th/viralpneumonia> และ API : <https://covid19.ddc.moph.go.th/>

ข้อมูลผู้ติดเชื้อร้ายแรงประจำวันที่ไม่ได้มาจากรายงานผลตรวจในเบื้องต้น เป็นรายงานควบคุมโรคต้องดำเนินการตรวจสอบข้อมูลในรายเดือนให้มีถูกต้องก่อนนำมารายงาน เพื่อใช้ในการวางแผนและประเมินผล

บุบ Wong :  

 Data API  ผังตัว  แสดงผลเต็ม...

ค้นหาข้อมูล... 

ข้อมูลทั้งหมด 192721 รายการ แสดงข้อมูลลำดับที่ 1 ถึง 100 

	_id	No.	announce_date	Notified date	sex	age
1	1	816990	2021-08-12T...	2021-08-11T0...	ชาย	7
2	2	816991	2021-08-12T...	2021-08-11T0...	ชาย	1
3	3	816992	2021-08-12T...	2021-08-11T0...	ชาย	35
4	4	816993	2021-08-12T...	2021-08-11T0...	หญิง	33
5	5	816994	2021-08-12T...	2021-08-11T0...	หญิง	14
6	6	816995	2021-08-12T...	2021-08-11T0...	ชาย	39
7	7	816996	2021-08-12T...	2021-08-11T0...	ชาย	45
8	8	816997	2021-08-12T...	2021-08-11T0...	หญิง	32
9	9	816998	2021-08-12T...	2021-08-11T0...	หญิง	7
10	10	816999	2021-08-12T...	2021-08-11T0...	ชาย	8
11	11	817000	2021-08-12T...	2021-08-11T0...	ชาย	27
12	12	817001	2021-08-12T...	2021-08-11T0...	หญิง	56
13	13	817002	2021-08-12T...	2021-08-11T0...	ชาย	22

<https://data.go.th/dataset/covid-19-daily>

Forms of tabular data

Master table

- Each row represent the unique unit of that table
- Usually relates to transactional table in 1-to-many forms
- Sometime, they are merged with transactional data (in bad database design / big data table)
- Example
 - Product tables
 - Customer tables
 - Machine description tables
 - Patient demographic data

Example: Customer table

Primary key identifies unique customer

Attributes of each customer



A	B	C	D	E	F	G	H	I	J	
1	CustomerKey	GeographyKey	Name	BirthDate	MaritalStatus	Gender	YearlyIncome	NumberChildrenAtHome	Occupation	HouseOwnerFlag
2	11000	26	Jon Yang	08/04/1986	M	M	90000		0 Professional	1
3	11001	37	Eugene Huang	14/05/1985	S	M	60000		3 Professional	0
4	11002	31	Ruben Torres	12/08/1985	M	M	60000		3 Professional	1
5	11003	11	Christy Zhu	15/02/1988	S	F	70000		0 Professional	0
6	11004	19	Elizabeth Johnson	08/08/1988	S	F	80000		5 Professional	1
7	11005	22	Julio Ruiz	05/08/1985	S	M	70000		0 Professional	1
8	11006	8	Janet Alvarez	06/12/1985	S	F	70000		0 Professional	1
9	11007	40	Marco Mehta	09/05/1984	M	M	60000		3 Professional	1
10	11008	32	Rob Verhoff	07/07/1984	S	F	60000		4 Professional	1
11	11009	25	Shannon Carlson	01/04/1984	S	M	70000		0 Professional	0
12	11010	22	Jacquelyn Suarez	06/02/1984	S	F	70000		0 Professional	0
13	11011	22	Curtis Lu	04/11/1983	M	M	60000		4 Professional	1
14	11012	611	Lauren Walker	18/01/1988	M	F	100000		0 Management	1
15	11013	543	Ian Jenkins	06/08/1988	M	M	100000		0 Management	1
16	11014	634	Sydney Bennett	09/05/1988	S	F	100000		0 Management	0
17	11015	301	Chloe Young	27/02/1999	S	F	30000		0 Skilled Manual	0
18	11016	329	Wyatt Hill	28/04/1999	M	M	30000		0 Skilled Manual	1

Example: Thailand province data

ชุดข้อมูล กลุ่มชุดข้อมูล

ชุดข้อมูลจังหวัดและภูมิภาคในประเทศไทย

องค์กร : สำนักงานพัฒนาธุร不做ดิจิทัล (องค์การมหาชน)

 ดาวน์โหลดทั้งหมด

ชุดข้อมูล "การแบ่งภูมิภาคและการแบ่งพื้นที่รับผิดชอบรายจังหวัดของหน่วยงานของรัฐระดับกรมหรือเทียบเท่าจำนวน 69 หน่วยงาน กว่า 90 แบบ ครอบคลุม 77 จังหวัดในประเทศไทย (รวมกรุงเทพมหานคร)" ข้อมูลชุดนี้รวมรวมและตรวจสอบจากแหล่งอ้างอิงที่ได้รับการยืนยัน คือ 1. กฎหมายจัดตั้งหน่วยงาน เริ่บใช้ตั้งแต่ปี พ.ศ. 2557 ถึงปัจจุบัน 2. สพร. โกรส์พาร์ก สถาบันเจ้าหน้าที่ของหน่วยงานเพื่อยืนยันความถูกต้องของข้อมูลอีกด้วย ไฟล์นี้มีวัตถุประสงค์เพื่อศึกษาและวิเคราะห์ข้อมูลทางภูมิศาสตร์ของประเทศไทย

 ดาวน์โหลด	ชุดข้อมูลจังหวัดและภูมิภาคในประเทศไทย  326 downloads
 ดาวน์โหลด	ชุดข้อมูลจังหวัดและภูมิภาคในประเทศไทย  108 downloads
 ดาวน์โหลด	ชุดข้อมูลจังหวัดและภูมิภาคในประเทศไทย  129 downloads
 ดาวน์โหลด	Data Dictionary ฉบับสมบูรณ์  66 downloads
 ดาวน์โหลด	Data Dictionary ฉบับสมบูรณ์  163 downloads

	_id	ProvinceNo	ProvinceMOI_ID	ProvinceNameT	ProvinceNameE	ProvinceBudget	F
1	1	1	10	กรุงเทพมหาน...	กท	75002	↑
2	2	2	81	จังหวัดกรุงบ...	กบ	70074	↑
3	3	3	71	จังหวัดกาญจน...	กจ	70041	↑
4	4	4	46	จังหวัดกาฬสินธ...	กส	70124	↑
5	5	5	62	จังหวัดกำแพง...	กพ	70181	↑
6	6	6	40	จังหวัดขอนแก่น	ขก	70122	↑
7	7	7	22	จังหวัดจันทบุรี	จบ	70091	↑

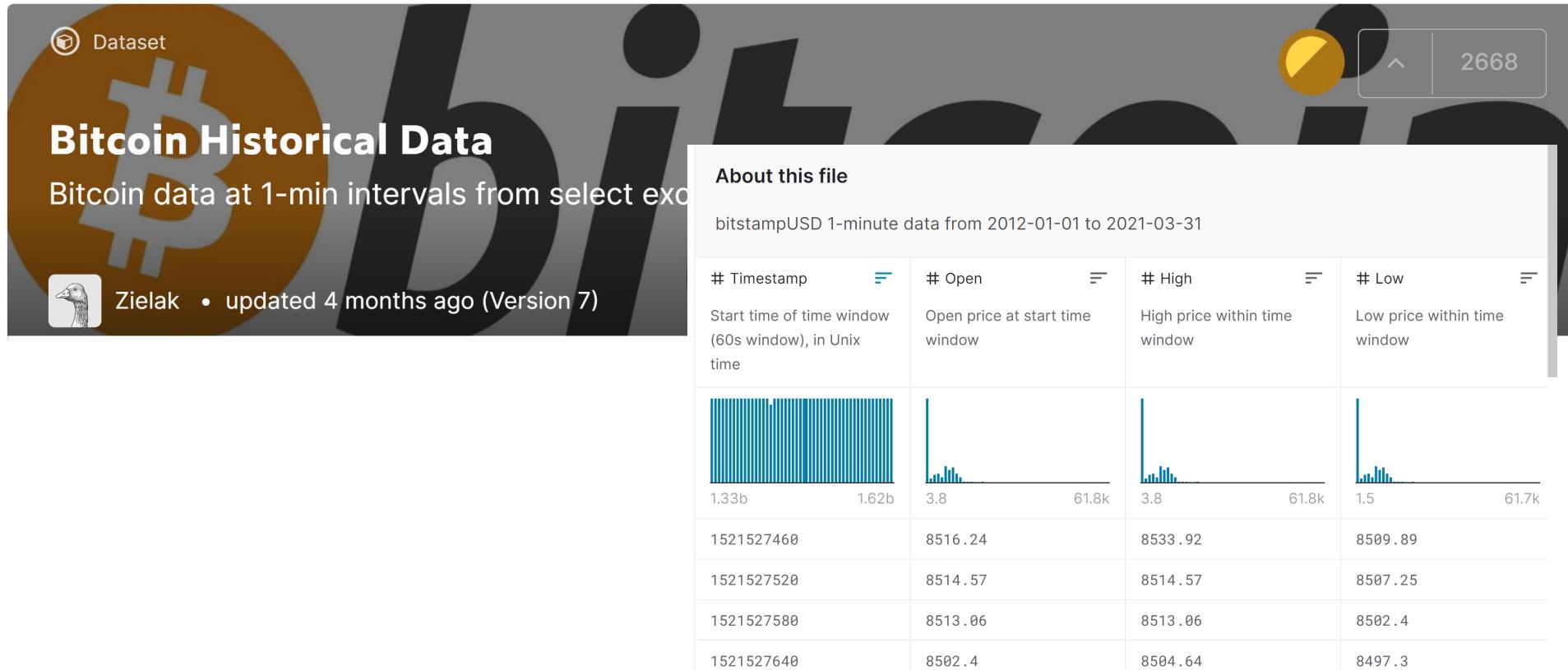
<https://data.go.th/dataset/proviceandregionthailand>

Forms of tabular data

Time series data

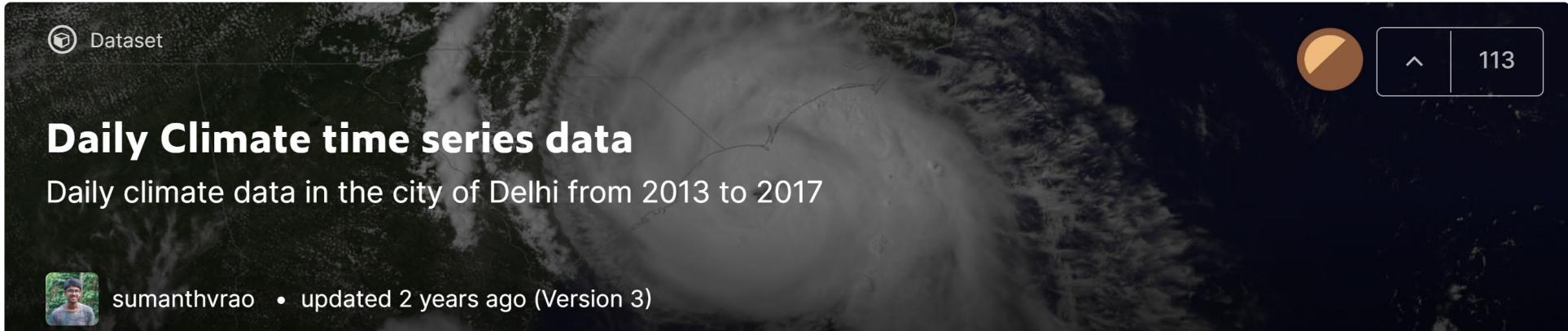
- Attributes of interest change by time
 - Also known as signals (more on signal data lecture)
- Often store each row for one time unit
- Need to prepare to suitable unit before use
- Example
 - Stock price
 - Machine status data
 - Weather data
 - Heart rate monitoring

Example: Bitcoin historical data



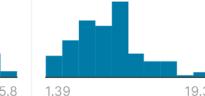
<https://www.kaggle.com/mczielinski/bitcoin-historical-data>

Example: climate data



About this file

This file contains weather data collected in the city of Delhi from the period of 4 years (from 2013 to 2017). It can be used for the purpose of training. This is purely academic dataset and is developed as a part of Data Analytics course of 2019 at PES University, Bangalore.

# date	# meantemp	# humidity	# wind_speed
Date of format YYYY-MM-DD	Mean temperature averaged out from multiple 3 hour intervals in a day.	Humidity value for the day (units are grams of water vapor per cubic meter volume of air).	Wind speed measured in kmph.
 1Jan17	 34.5	 95.8	 19.3
2017-01-01	15.91304347826087	85.8695652173913	2.743478260869565
2017-01-02	18.5	77.2222222222223	2.894444444444444
2017-01-03	17.11111111111111	81.8888888888889	4.0166666666666667
2017-01-04	18.7	70.05	4.545
2017-01-05	18.3888888888889	74.94444444444444	3.3000000000000003

<https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data>

Forms of tabular data

Graph/network data

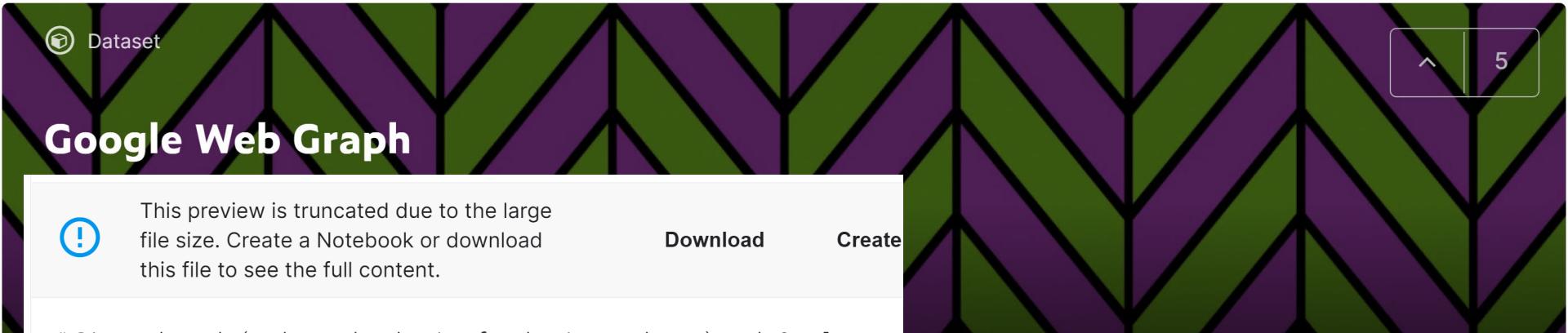
- Graph contains nodes/vertices and edges/connections
- Can have list or matrix forms
- List form
 - At least two columns (source, destination) representing two nodes
 - Each row represent each edge
 - May have other attributes
- Matrix form
 - Rarely used, bad design, very sparse
 - Row and column represent nodes while the value mark the existence or weights of the edges

Example: movies rating

	User id	Movie id	Rating	Timestamp
1	196	242	3	881250949
2	186	302	3	891717742
3	22	377	1	878887116
4	244	51	2	880606923
5	166	346	1	886397596
6	298	474	4	884182806
7	115	265	2	881171488
8	253	465	5	891628467
9	305	451	3	886324817
L0	6	86	3	883603013
L1	62	257	2	879372434
L2	286	1014	5	879781125
L3	200	222	5	876042340
L4	210	40	3	891035994
L5	224	29	3	888104457
L6	303	785	3	879485318
L7	122	387	5	879270459
L8	194	274	2	879539794
L9	291	1042	4	874834944
L0	234	1184	2	892079237

Each movie review

Example: Google web hyperlinks



<https://www.kaggle.com/pappukrjha/google-web-graph>

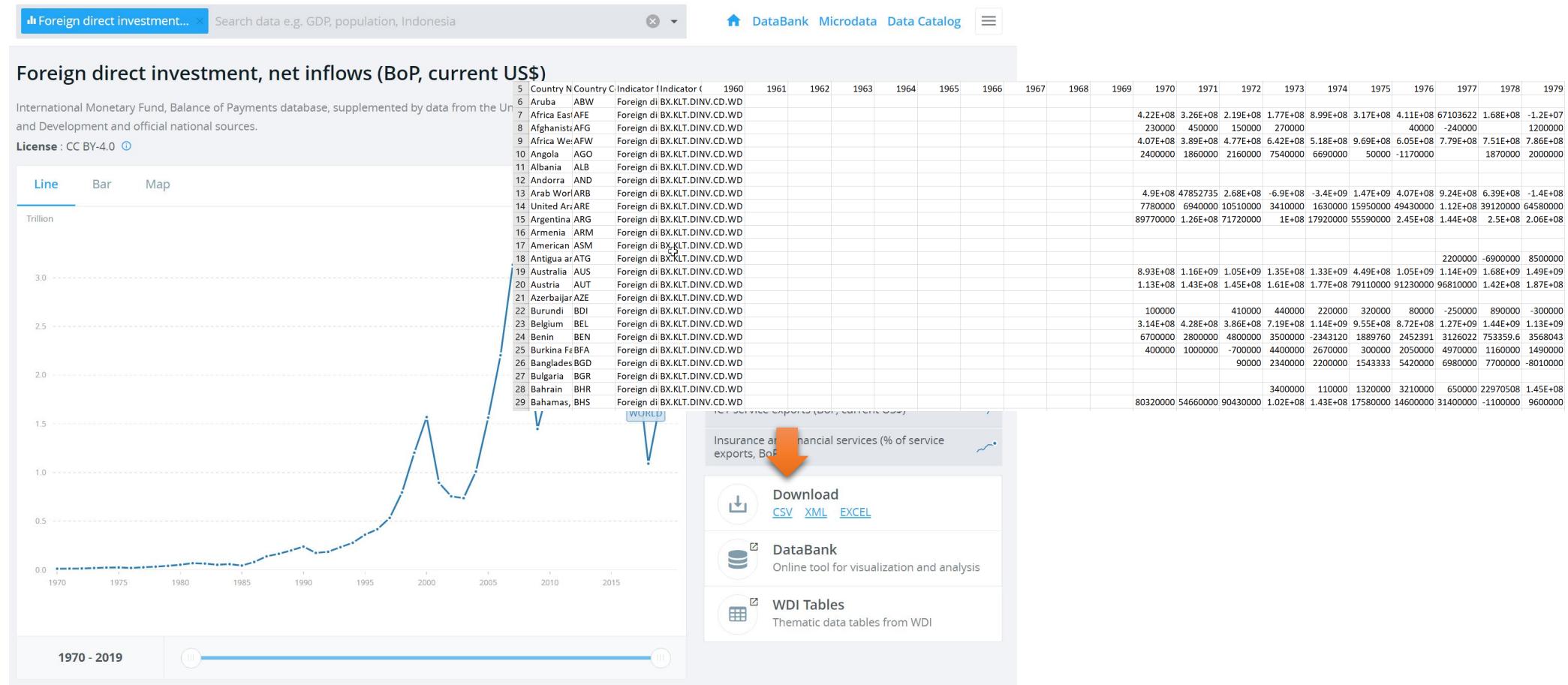
Forms of tabular data

Crosstable data

- Row and column represent attributes (one of them are often growing by time/case)
- Use for displaying the multiple summarized attributes change by time
- Often found in summary reports
- Examples
 - Economic report
 - World Bank report
 - Government report



Example: Foreign direct investment, net inflow



Example: Income/expense by family

รายได้ของครัวเรือน

องค์กร : สำนักงานสต๊อกแห่งชาติ

https://data.go.th/dataset/ns_08_20241

โครงการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือน รายได้ของครัวเรือน นายสิงห์ "เงินหรือสีงของ" ที่ครัวเรือนได้รับมาจากการทำงานหรือผลิตเอง หรือจากทรัพย์สินหรือได้รับความช่วยเหลือจากผู้อื่น จำแนกเป็น รายได้ประจำ และรายได้ไม่ประจำ

ดาวน์โหลด	รายได้เฉลี่ยต่อเดือนของครัวเรือน จำนวนรายได้  2,992 downloads
ดาวน์โหลด	รายได้เฉลี่ยต่อเดือนของครัวเรือน จำนวนรายได้ ...  1,307 downloads
ดาวน์โหลด	ร้อยละของครัวเรือน จำนวนรายได้ก้าวสีน้ำเงินเฉลี่ยต่อเดือน  631 downloads
ดาวน์โหลด	ร้อยละของครัวเรือน จำนวนรายได้ก้าวสีน้ำเงินเฉลี่ยต่อเดือน ...  266 downloads
ดาวน์โหลด	ส่วนแบ่งรายได้ประจำของครัวเรือน โดยการจำแนกครัวเรือนเป็น 5 กลุ่ม  276 downloads
ดาวน์โหลด	รายได้เฉลี่ยต่อเดือนของครัวเรือน จำนวนรายได้  926 downloads

บุบบอง :   Data API Visualization ผังตัว แสดงผลเต็ม...

ค้นหาข้อมูล... 

ข้อมูลทั้งหมด 108 รายการ แสดงข้อมูลลำดับที่ 1 ถึง 100 >

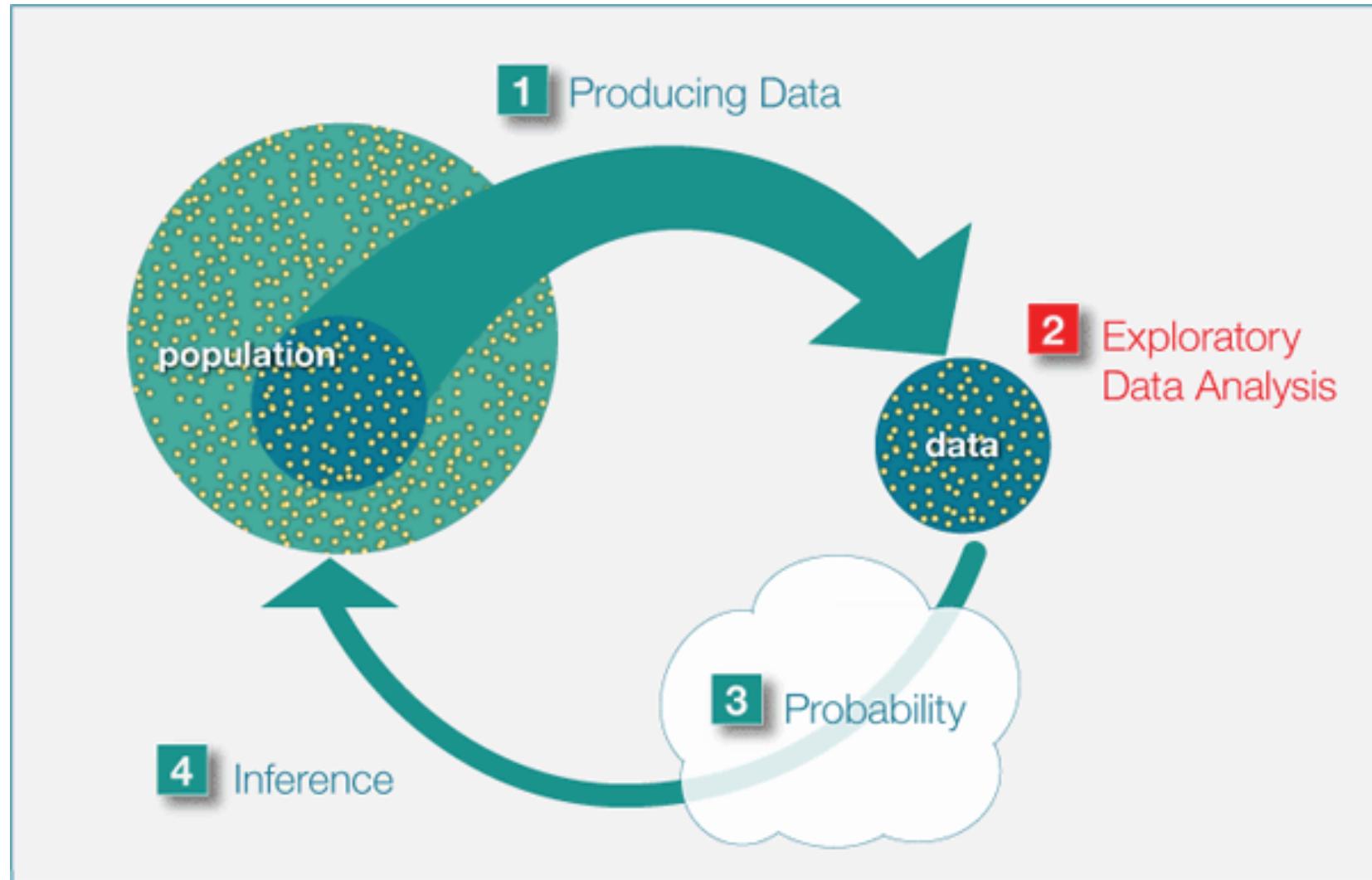
序	_id	YEAR	CODE_REGION	REGION	CODE_AREA	AREA	MO
01	01	2562	2	กลาง	2	นอกราชอาณาจ	24%
81	81	2562	3	เมือง	0	นอกราชอาณาจ	24%
82	82	2562	3	เมือง	1	ในเขตเทศบาล	20%
83	83	2562	3	เมือง	2	ในเขตเทศบาล	183
84	84	2562	4	ตะวันออกเฉียง...	0	เขตการปกครอง	20%
85	85	2562	4	ตะวันออกเฉียง...	1	ในเขตเทศบาล	257
86	86	2562	4	ตะวันออกเฉียง...	2	นอกราชอาณาจ	182
87	87	2562	5	ใต้	0	เขตการปกครอง	25%
88	88	2562	5	ใต้	1	ในเขตเทศบาล	26%
89	89	2562	5	ใต้	2	นอกราชอาณาจ	24%
90	90	2564	0	ก้าวราชอาณาจ...	0	เขตการปกครอง	27%
91	91	2564	0	ก้าวราชอาณาจ...	1	ในเขตเทศบาล	317
92	92	2564	0	ก้าวราชอาณาจ...	2	นอกราชอาณาจ...	23%
93	93	2564	0	ก้าวราชอาณาจ...	3	ในเขตเทศบาล	24%

Exploratory Data Analysis

- Exploratory data analysis or “EDA” is a critical step in analyzing the data
- The main reasons are
 - detection of mistakes, outliers or abnormalities
 - checking of assumptions
 - preliminary selection of appropriate models
 - determining relationships among the explanatory variables
 - assessing the direction and rough size of relationships between explanatory and outcome variables



Exploratory Data Analysis



1. Load data and preprocess data

```
telcoData.isnull().any()
```

```
telcoData.head()
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	
	5575-GNVDE	Male	0	No	No	34	Yes	No	
	3668-QPYBK	Male	0	No	No	2	Yes	No	
	7795-CFOCW	Male	0	No	No	45	No	No phone service	
	9237-HQITU	Female	0	No	No	2	Yes	No	

gender	False
SeniorCitizen	False
Partner	False
Dependents	False
tenure	False
PhoneService	False
MultipleLines	False
InternetService	False
OnlineSecurity	False
OnlineBackup	False
DeviceProtection	False
TechSupport	False
StreamingTV	False
StreamingMovies	False
Contract	False
PaperlessBilling	False
PaymentMethod	False
MonthlyCharges	False
TotalCharges	False
Churn	False
dtype:	bool

Univariate analysis

Categorical (factors/strings)

- Use frequency table
- Inspect the prior probabilities, aka proportions
- Identify types of categories: nominal vs ordinal
- Identify the ID columns. These columns will need to be removed before analysis. Why?
- Take note at minority categories

Frequency table: one-variable

```
pd.crosstab(telcoData[ 'Churn' ], columns='Count')
```

col_0 Count

Churn

No 5174

Yes 1869

```
pd.crosstab(telcoData[ 'Churn' ], columns='Count', normalize='columns')
```

col_0 Count

Churn

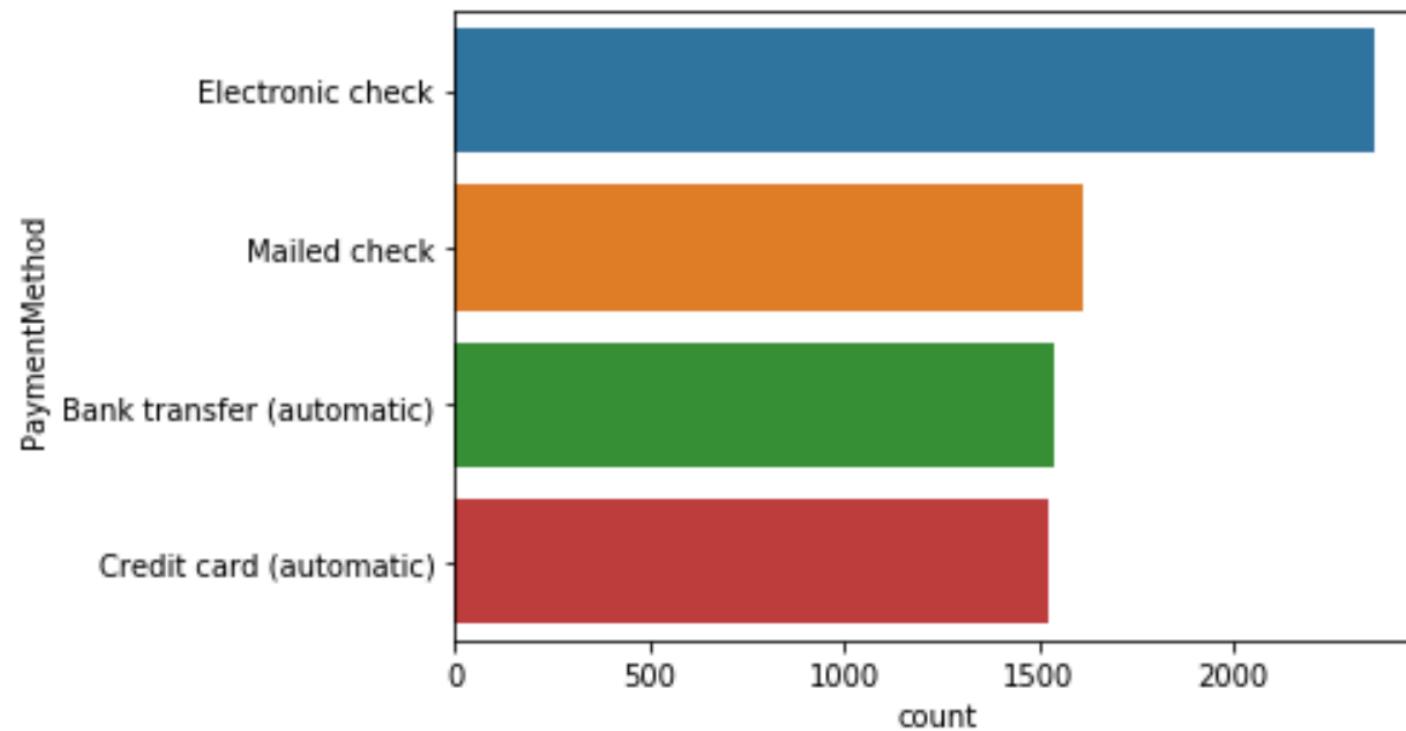
No 0.73463

Yes 0.26537

Countplot

```
import seaborn as sns
sns.countplot(data = telcoData, y = 'PaymentMethod')
```

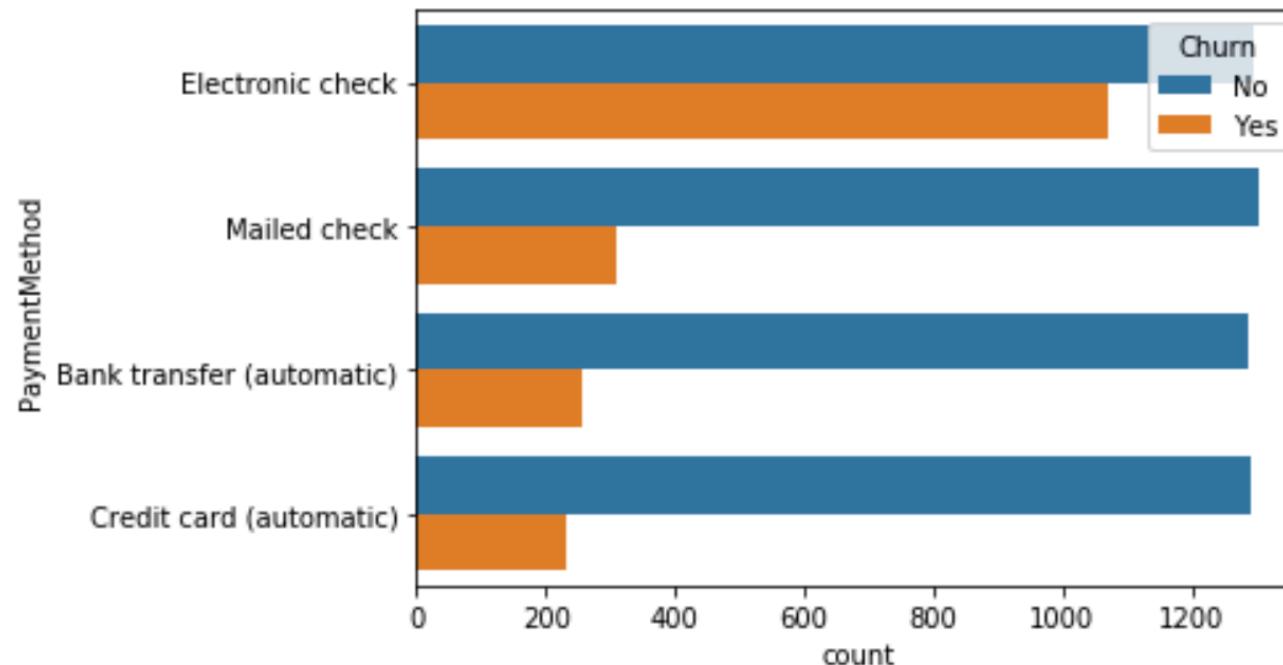
```
<matplotlib.axes._subplots.AxesSubplot at 0x12264ff60>
```



Countplot with multiple x variables

```
import seaborn as sns
sns.countplot(data = telcoData, y = 'PaymentMethod', hue = 'Churn')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x123f286a0>
```



Univariate analysis

Numeric

- Check distribution
- Center, spread, skewness, outliers
- Identify numeric columns that should actually be converted into categorical
- Use histogram to inspect the distribution

Descriptive statistics: describe

```
telcoData.describe()
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.734304
std	0.368612	24.559481	30.090047	2266.794470
min	0.000000	0.000000	18.250000	0.000000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

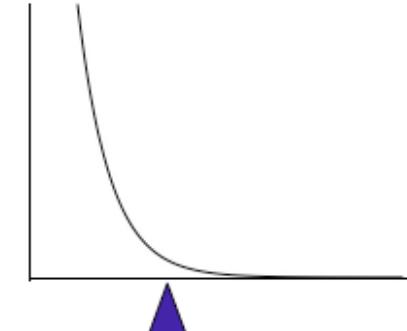
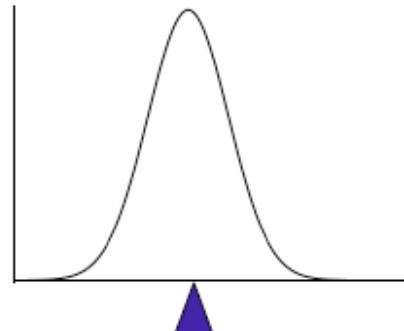
Central tendency

Mean

I. The Mean

To calculate the average \bar{x} of a set of observations, add their value and divide by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$



Central tendency

Median

- **Median** – the exact middle value
- **Calculation:**
 - If there are an odd number of observations, find the middle value
 - If there are an even number of observations, find the middle two values and average them
- **Example**

Some data:

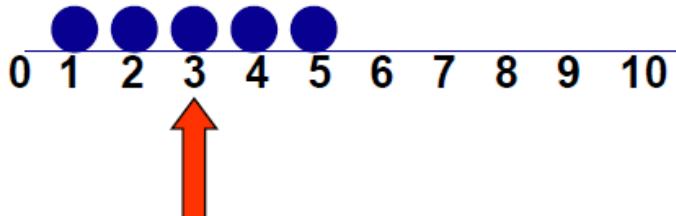
Age of participants: 17 19 21 22 23 23 23 38

$$\text{Median} = (22+23)/2 = 22.5$$

Central tendency

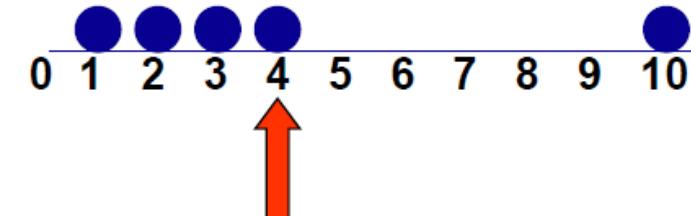
Which location measure is the best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers



Mean = 3

Median = 3



Mean = 4

Median = 3

Scale: Variance

- Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Why squared deviations?

- Adding deviations will yield a sum of ?
- Absolute values do not have nice mathematical properties (non-linear)
- Squares eliminate the negatives
- Results:
 - Increasing contribution to the variance as you go farther from the mean

Scale: Variance

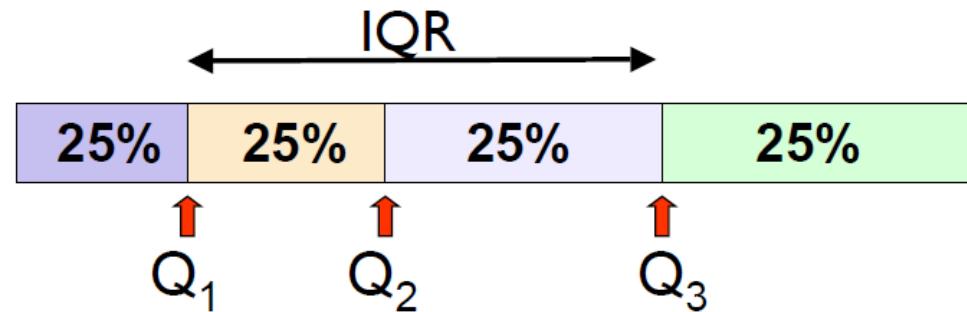
- Variance is somewhat arbitrary
- What does it mean to have a variance of 8.9? Or 1.5? Or 1245.34? Or 0.00001?
- Nothing. But if you could “standardize” that value, you could talk or compare about any variance (i.e. deviation) in equivalent terms
- Standard deviations are simply the square root of the variance

Scale: Standard deviation

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

1. Score (in the units that are meaningful)
2. Mean
3. Each score's deviation from the mean
4. Square that deviation
5. Sum all the squared deviations (Sum of Squares)
6. Divide by $n-1$
7. Square root – now the value is in the units we started with!!!

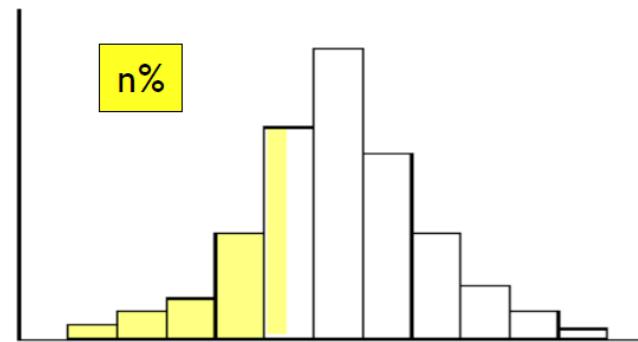
Scale: Quartiles and IQR



- The first quartile, Q_1 , is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Percentiles (aka Quantiles)

In general the **nth percentile** is a value such that n% of the observations fall at or below or it



$Q_1 = 25^{\text{th}}$ percentile

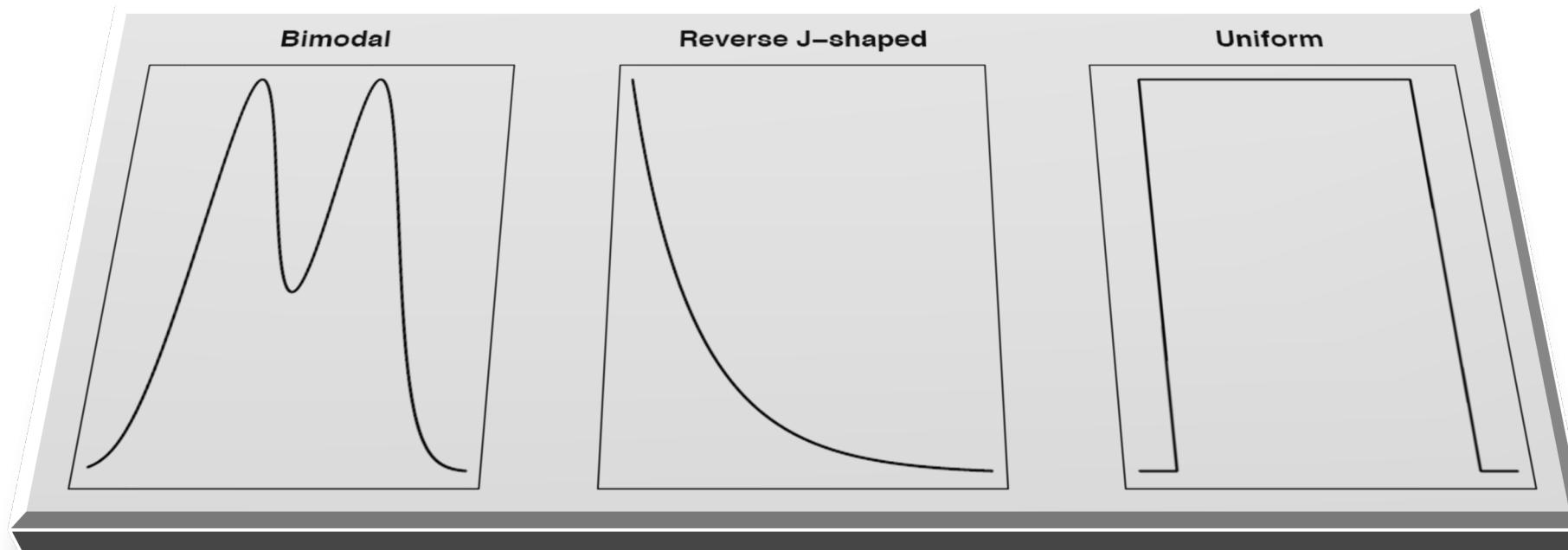
Median = 50^{th} percentile

$Q_2 = 75^{\text{th}}$ percentile

Common distribution shapes



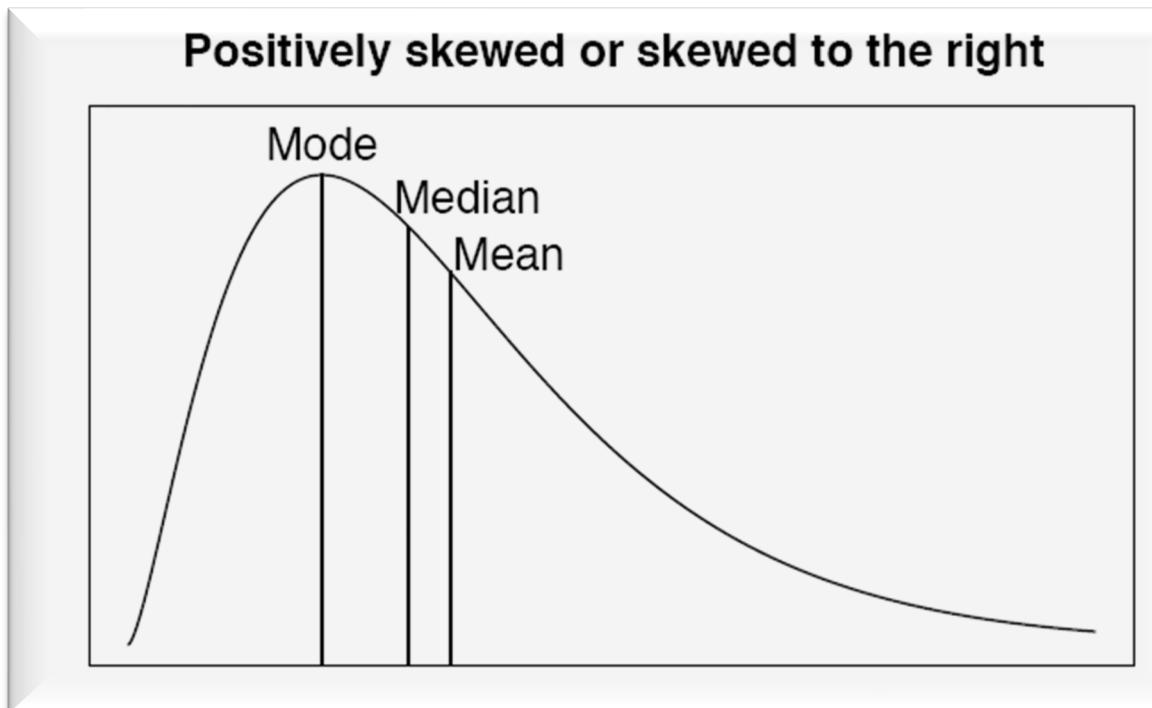
Other distribution shapes



Skewness I

Positively skewed

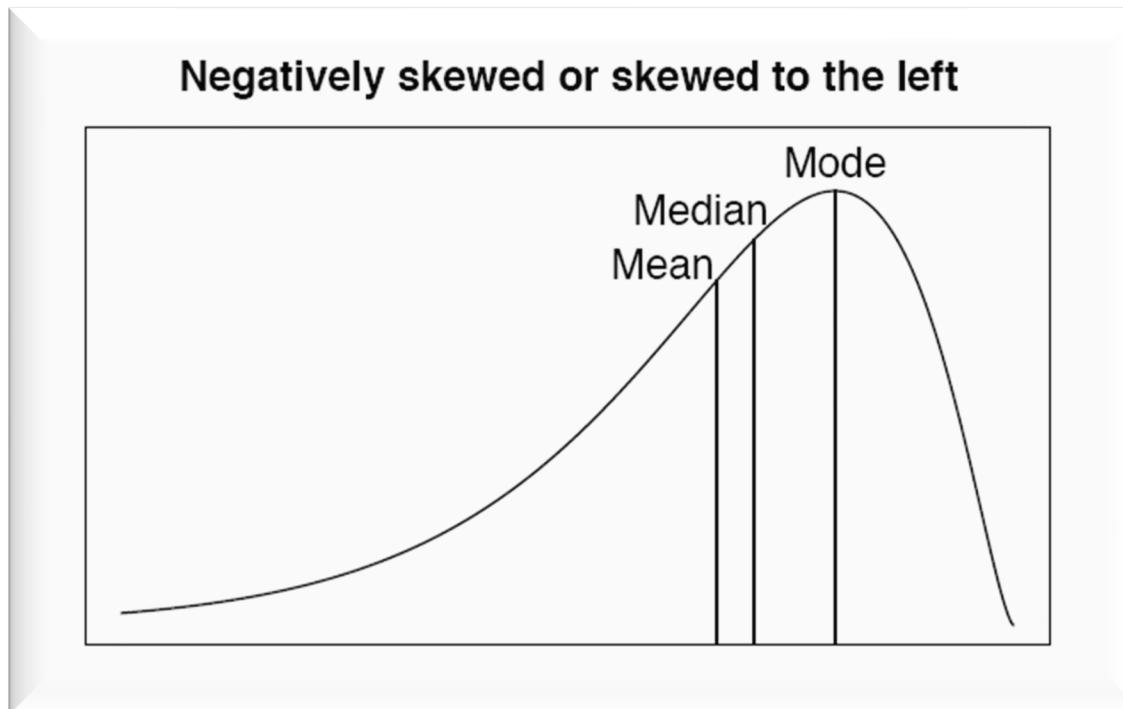
- Longer tail in the high value
- Mean > Median > Mode



Skewness II

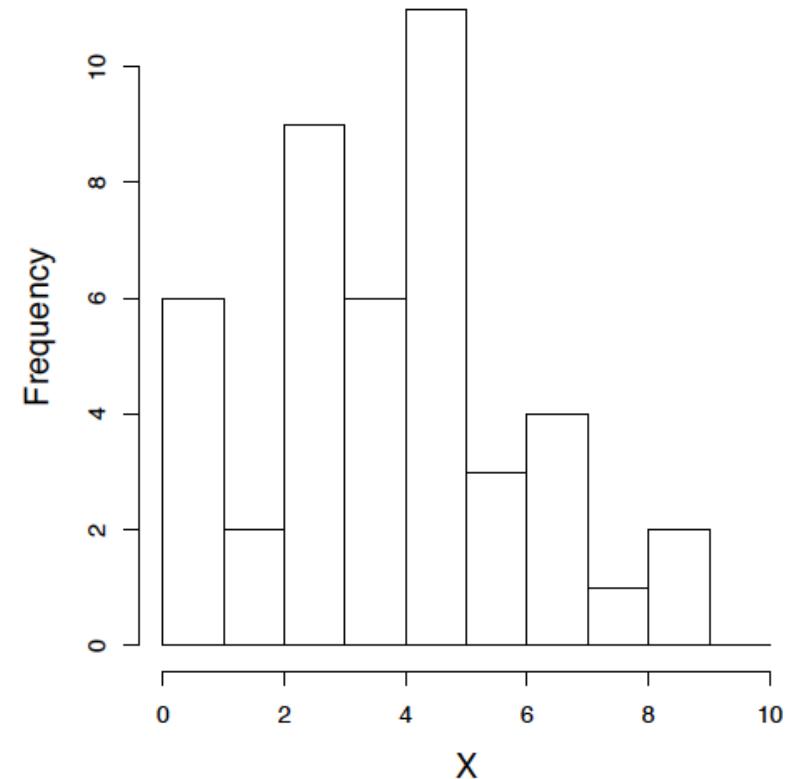
Negatively skewed

- Longer tail in the low value
- Mode > Median > Mean



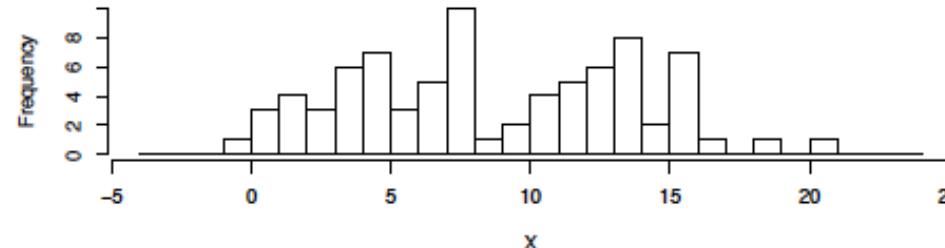
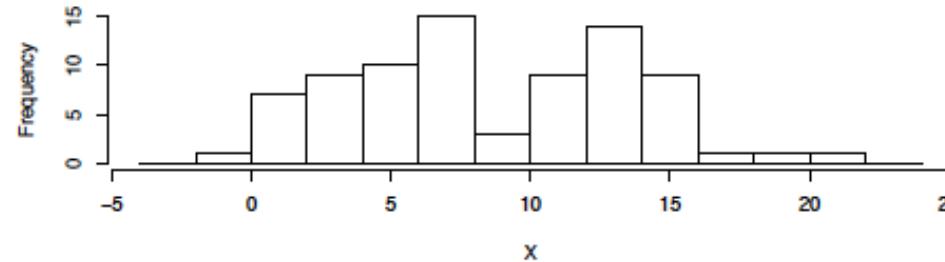
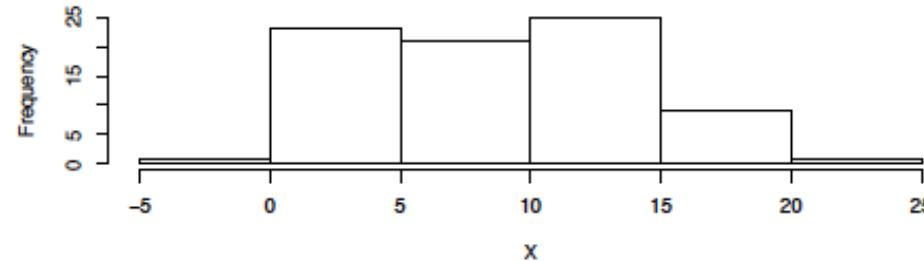
Univariate Graphical EDA Histogram

- Histogram is a graphical representation of the distribution of numerical data
- It provides a view of data density and the shape of data distribution
- To construct a histogram, the first step is to
 - bin the range of values
 - count how many values fall into each interval
- The bins are usually specified as consecutive, non-overlapping intervals of a variable.
- The bins (intervals) must be adjacent, and are usually equal size.



Univariate Graphical EDA

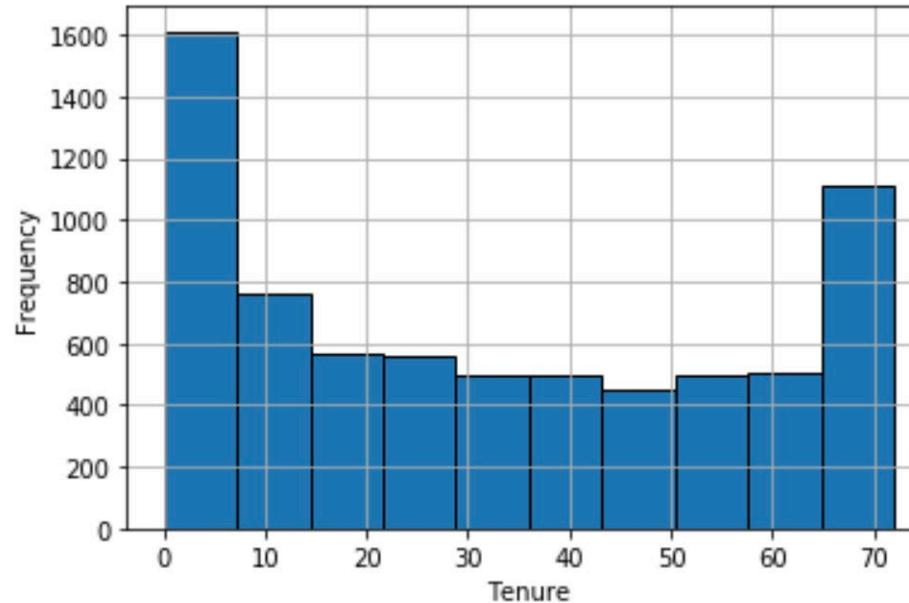
Effects of Histogram Bin



Histogram with Pandas: tenure

```
import matplotlib.pyplot as plt
telcoData['tenure'].hist(edgecolor='black')
plt.xlabel('Tenure')
plt.ylabel('Frequency')

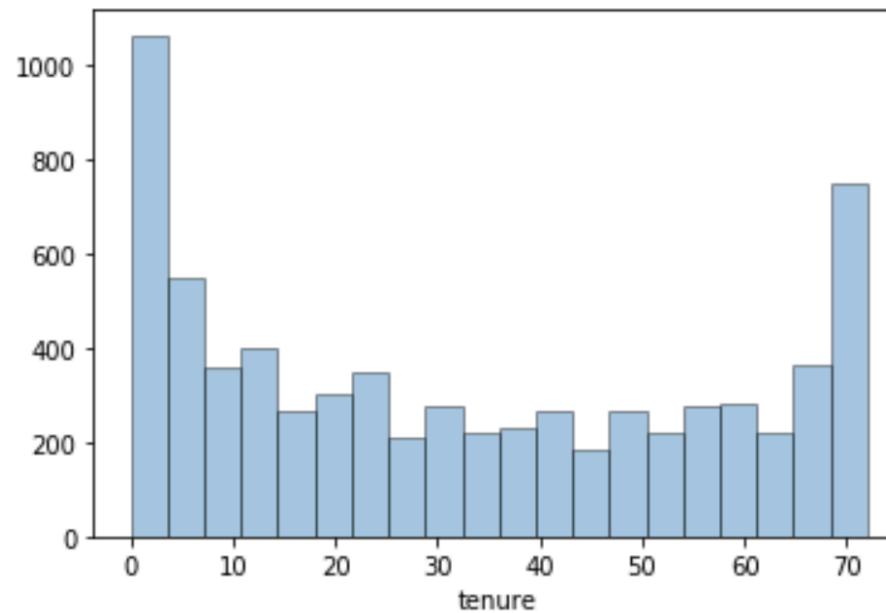
Text(0, 0.5, 'Frequency')
```



Histogram with seaborn: tenure

```
sns.distplot(telcoData[ 'tenure' ],
             bins=20,
             kde=False,
             hist_kws={ 'edgecolor':'black'})
```

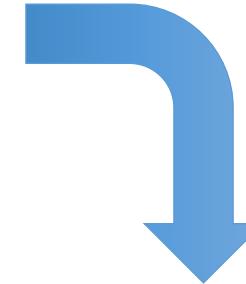
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2b30dc50>
```



Multivariate analysis

Categorical vs Categorical

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M



Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Pandas: Crosstab

```
pd.crosstab( index = telcoData['Contract'],
              columns = telcoData['Churn'])
```

	Churn	No	Yes
Contract			
Month-to-month	2220	1655	
One year	1307	166	
Two year	1647	48	

Pandas: normalized crosstab

```
pd.crosstab( index = telcoData['Contract'],
              columns = telcoData['Churn'],
              normalize=True)
```

	Churn	No	Yes
Contract			
Month-to-month	0.315207	0.234985	
One year	0.185574	0.023570	
Two year	0.233849	0.006815	

High Risk group

Low Risk group

Pandas: pivot_table

```
import numpy as np
pd.pivot_table(data = telcoData,
                index = ['Contract','PaymentMethod'],
                columns = 'Churn',
                values = 'TotalCharges',
                aggfunc= np.size) \
    .apply(lambda x: x/sum(x), axis = 1)
```

		Churn	No	Yes
		Contract	PaymentMethod	
Month-to-month	Bank transfer (automatic)	0.658744	0.341256	
	Credit card (automatic)	0.672192	0.327808	
	Electronic check	0.462703	0.537297	
	Mailed check	0.684211	0.315789	
	One year	Bank transfer (automatic)	0.902813	0.097187
		Credit card (automatic)	0.896985	0.103015
One year		Electronic check	0.815562	0.184438
		Mailed check	0.931751	0.068249

Multivariate analysis

Categorical vs Numerical

Graphical

- Nominal vs Numeric: Bar Chart
- Ordinal vs Numeric: Bar/Line Chart
- Categorical vs Distribution: Histogram + Styling

Non-Graphical

- dplyr: group_by + summarise
- statistics: ANOVA, t-test, Spearman's rank regression

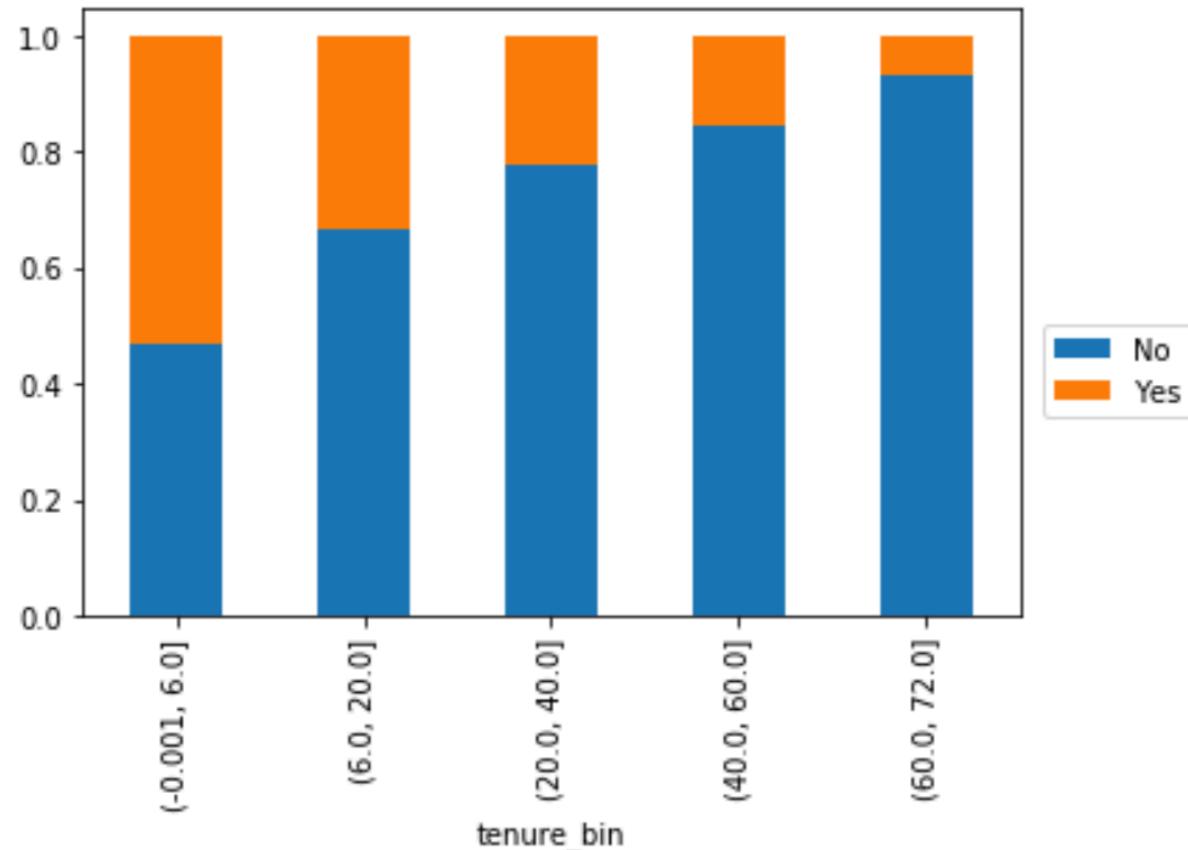
What to look for?

- Relationship to target variable
 - Linear
 - Proportion
 - Specificity
- Dependency
 - Change in distribution when one variable changes

```
telcoData['tenure_bin'] = pd.qcut(telcoData['tenure'], q = 5)
```

```
churnByTenure = pd.crosstab(index = telcoData['tenure_bin'],
                             columns = telcoData['Churn'])\n                             .apply(lambda x: x/sum(x), axis = 1)\nchurnByTenure.plot.bar(stacked = True).legend(bbox_to_anchor=(1.2, 0.5))
```

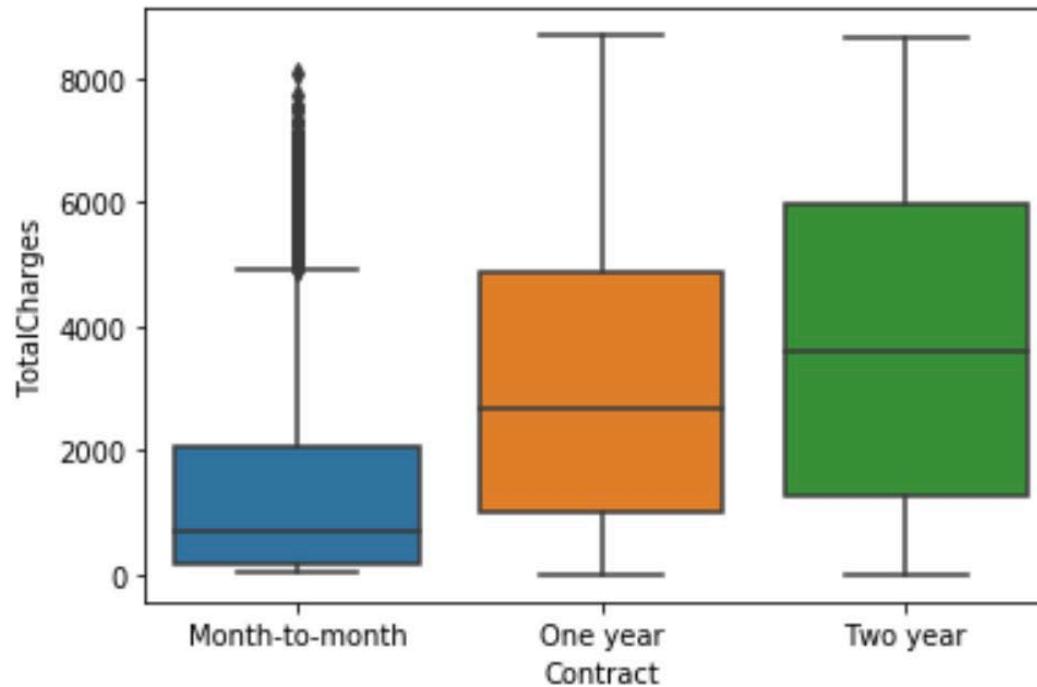
```
<matplotlib.legend.Legend at 0x1a2bc0beb8>
```



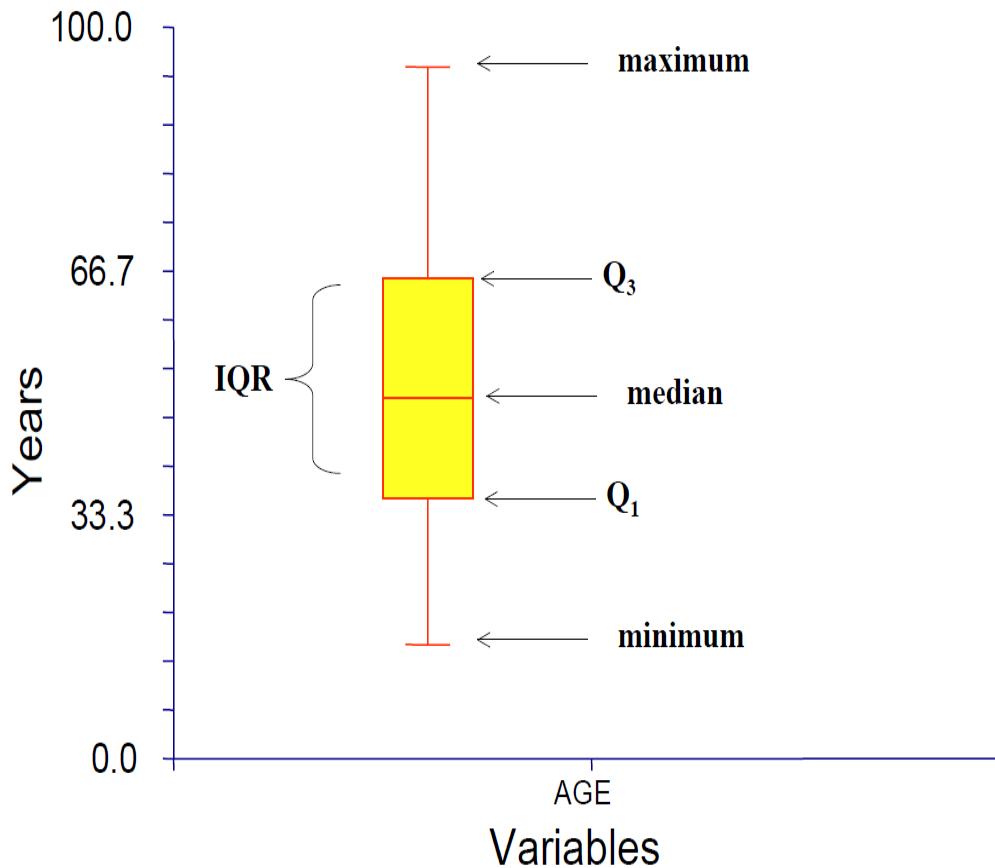
Boxplot

```
sns.boxplot(data = telcoData, x = 'Contract', y = 'TotalCharges')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2be3ec50>
```



Boxplot



- The box in boxplot represents the middle 50% of the data
- The middle line indicates median
- Whiskers can be designated as either
 - Max/Min
 - Outlier boundaries
 - Upper = $Q_3 + 1.5 * IQR$
 - Lower = $Q_1 - 1.5 * IQR$

Multivariate analysis

Numerical vs Numerical

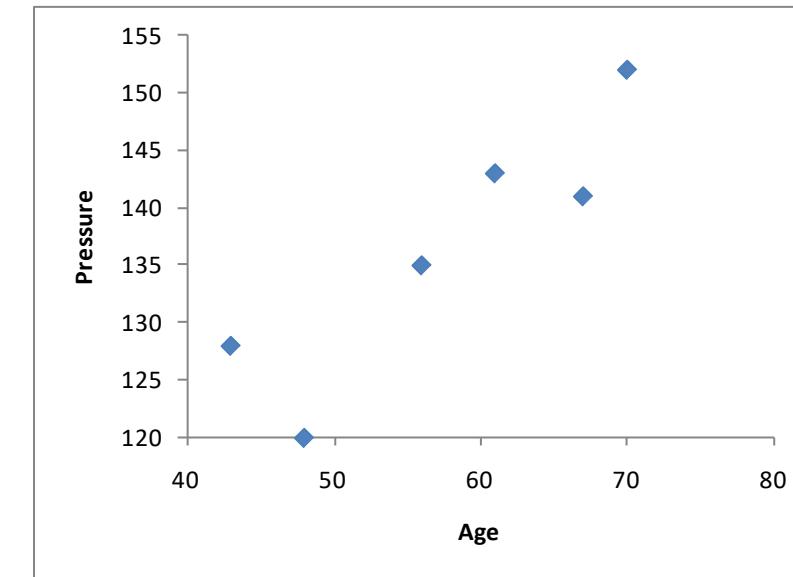
- Graphical
 - Scatter plot
- Non-graphical
 - Correlation

Multivariate Graphical EDA

Scatter plot

- A **scatter plot** is a graph of the ordered pairs (x,y) of numbers consisting of the independent variable x and the dependent variable y .

Subject	Age x	Pressure y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

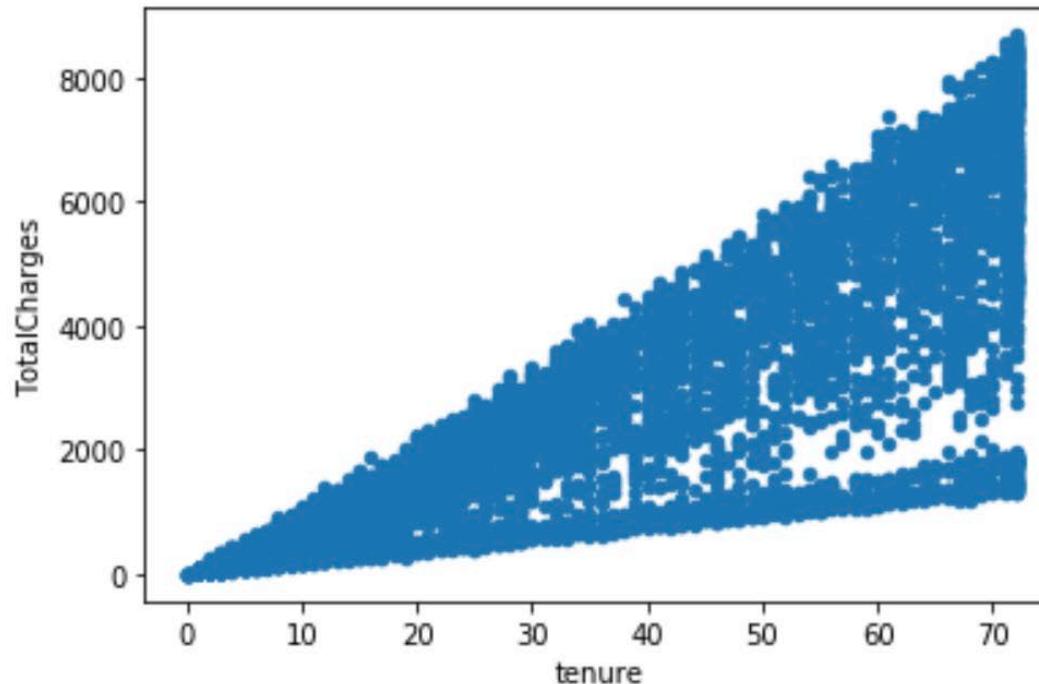


Scatter plot

Non-aggregated

```
telcoData.plot.scatter(x = 'tenure', y = 'TotalCharges')
```

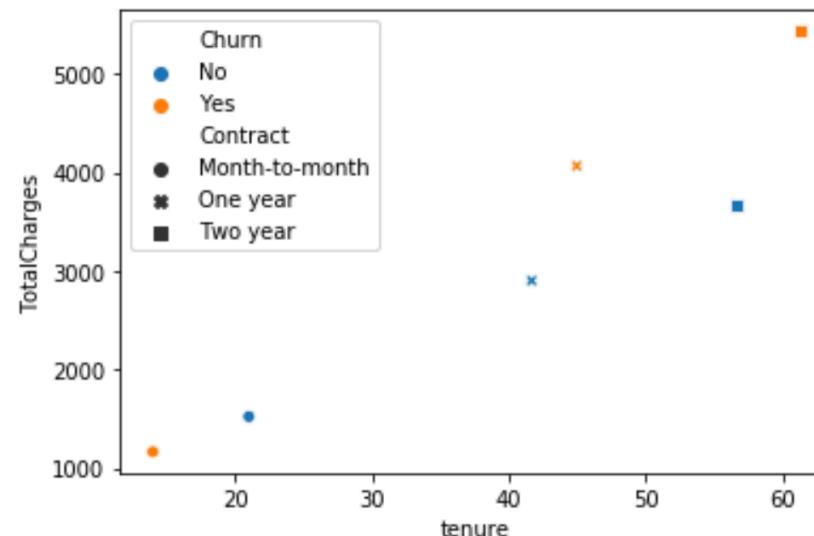
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2dc03c50>
```



Scatter plot Aggregated

```
telcoData1 = telcoData.groupby(['Contract', 'Churn'], as_index=False)\n                .agg({'tenure': 'mean', 'TotalCharges': 'mean'})\nsns.scatterplot(data = telcoData1,\n                  x = 'tenure',\n                  y = 'TotalCharges',\n                  hue = 'Churn',\n                  style = 'Contract')
```

<matplotlib.axes._subplots.AxesSubplot at 0x1a2dfbf98>



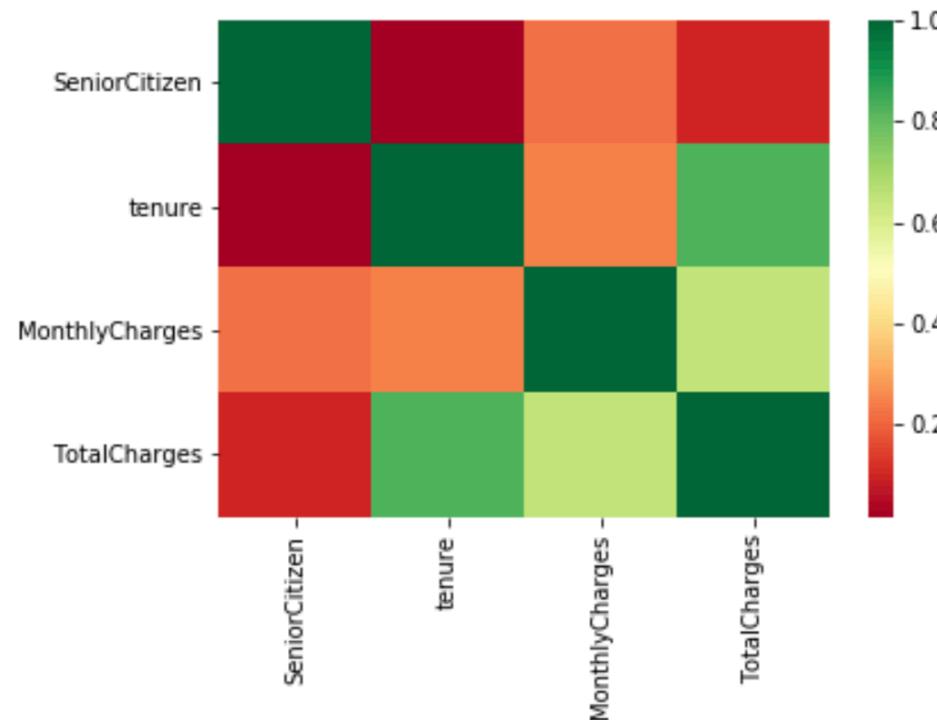
Correlation

```
telcoData.corr()
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
SeniorCitizen	1.000000	0.016567	0.220173	0.103006
tenure	0.016567	1.000000	0.247900	0.826178
MonthlyCharges	0.220173	0.247900	1.000000	0.651174
TotalCharges	0.103006	0.826178	0.651174	1.000000

Heatmap of correlation

```
sns.heatmap(telcoData.corr(), cmap = 'RdYlGn')  
<matplotlib.axes._subplots.AxesSubplot at 0x1a2f962048>
```



Lab

- Select a data source
- Identify what form of data it has
- Select one aspect/question of the data
- Create a visualization to describe the data



Thank you

Question?

