BIG DATA
EXPERIENCE

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

BIG DATA
EXPERIENCE
CENTER

KMUTT

BIG DATA EXPERIENCE

Topics

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

https://colab.research.google.com/drive/1Jp_etz6rejxOHl0lkgSK66A1VBXEneO-?usp=sharing

KMUTT

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 Cod	Market
2	32298	CA-2012-124	31/07/2012	31/07/2012	Same Day	RH-19495	Rick Hansen	Consumer	New York Cit	New York	United States	10024	US
3	26341	IN-2013-7787	05/02/2013	07/02/2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South W	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 M	Home Office	Berlin	Berlin	Germany		EU
6	47221	SG-2013-432	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 W	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 Cla	MB-18085	Mick Brown	Consumer	Hamilton	Waikato	New Zealand		APAC
10	40155	CA-2014-135	14/10/2014	21/10/2014	Standard Cla	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 Carolir	United States	28027	US
12	34577	CA-2011-102	05/04/2011	09/04/2011	Second Class	GM-14695	Greg Maxwel	Corporate	Alexandria	Virginia	United States	22304	US
13	28879	ID-2012-2840	19/04/2012	22/04/2012	First Class	AJ-10780	Anthony Jacc	Corporate	Kabul	Kabul	Afghanistan		APAC
14	45794	SA-2011-1836	27/12/2011	29/12/2011	Second Class	MM-7260	Magdelene M	Consumer	Jizan	Jizan	Saudi Arabia		EMEA
15	4132	MX-2012-130	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	Ile-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-2014-1651	08/09/2014	14/09/2014	Standard Cla	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-2014-8190	05/12/2014	07/12/2014	Second Class	RH-9555	Ritsa Hightov	Consumer	Uvinza	Kigoma	Tanzania		Africa
22	46630	PL-2012-7820	08/08/2012	10/08/2012	First Class	AB-600	Ann Blume	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-2011-4480	02/05/2011	03/05/2011	First Class	JK-15325	Jason Klamcz	Corporate	Suzhou	Anhui	China		APAC
25	13528	ES-2013-2860	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	Valerie Domi	Consumer	Soyapango	San Salvador	El Salvador		LATAM
28	30191	IN-2011-1028	17/12/2011	20/12/2011	First Class	PB-19210	Phillip Breyer	Corporate	Taipei	Taipei City	Taiwan		APAC

Format

- Database

```
1  SELECT * FROM SALES LIMIT 100;
```

SALES (100r × 10c)

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	219.58199999999997
CA-2013-138688	2014-06-13	2014-06-17	Second Class	DV-13045	OFF-LA-10000240	14.62	2	0.00	6.8713999999999995
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-Ca
tegory|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|Sout
h|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|Sout
h|FUR-CH-10000454|Furniture|Chairs|Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back|731.9399999999999|3|0|219.58199999999997
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|9
0036|West|OFF-LA-10000240|Office Supplies|Labels|Self-Adhesive Address Labels for Typewriters by Universal|14.62|2|0|6.8713999999999995
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|-383.03100000000006
```

- API (may need transformation)

```
{
    "success": true,
    "timestamp": 1620386223,
    "source": "GBP",
    "quotes": {
        "GBPAED": 5.109727,
        "GBPAFN": 107.750768,
        "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

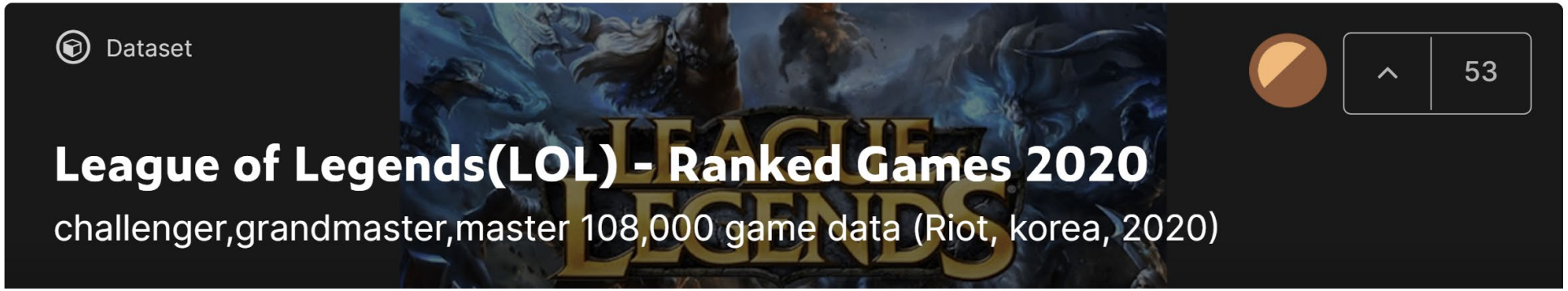
Example: sales transactions

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	Rounded Back	731.9399999999999	3	0	219.58199999999997														
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	6.8713999999999995
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	-383.03100000000006
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	22.368000000000002	2	0.2	2.516399999999999
6	CA-2011	Cherry Wood	48.86	7	0	14.169399999999996														
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	4	0	1.9656000000000002
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-10002275	Technology	Phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.71520000000004
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
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
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	1706.1840000000002	9	0.2	85.3091999999998
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 phoneÂ - Charcoal black	911.424	4	0.2	68.35680000000002
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	15.552000000000003	3	0.2	5.4432
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	407.97600000000006	3	0.2	132.5921999999999
15	US-201	Very Larg	HEPA Filter	68.80999999999999	5	0.8	-123.858													
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	-3.8160000000
17	CA-201	Vertical 5-Shelf: 72""H x 36""W x 18 1/2""D"	665.88	6	0	13.317599999999999														
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
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-10003056	Office Supplies	Art	Newell 341	8.56	2	0	2.4823999999999993
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	16.01099999999998
21	CA-201	White	1"""	22.72	4	0.2	7.383999999999999													
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	19.4599999999999997	7	0	5.0596
23	CA-201	4' Cord Length	60.3399999999999996	7	0	15.688400000000001														
24	US-201	Gray	71.37199999999999	2	0.3	-1.0196000000000005														
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	1044.6299999999999	3	0	240.2649
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
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 2.0Â Flash Drive	90.57000000000001	3	0	11.774100000
28	US-201	Royale Cherry Finish	3083.4300000000003	7	0.5	-1665.0522														
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	9.618000000000002	2	0.7	-7.0532
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"	124.200000000			
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
32	US-201	Putty/Woodgrain	86.304	6	0.2	9.709199999999989														
33	US-201	14 7/8"" x	Executive Red"	6.8580000000000001	6	0.7	-5.715													
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
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
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
37	CA-201	Black	190.92	5	0.6	-147.96300000000002														
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
39	CA-201	Custom Colors	532.3992	3	0.32	-46.97640000000001														

2-1-sales_transactions

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

Example: game transaction



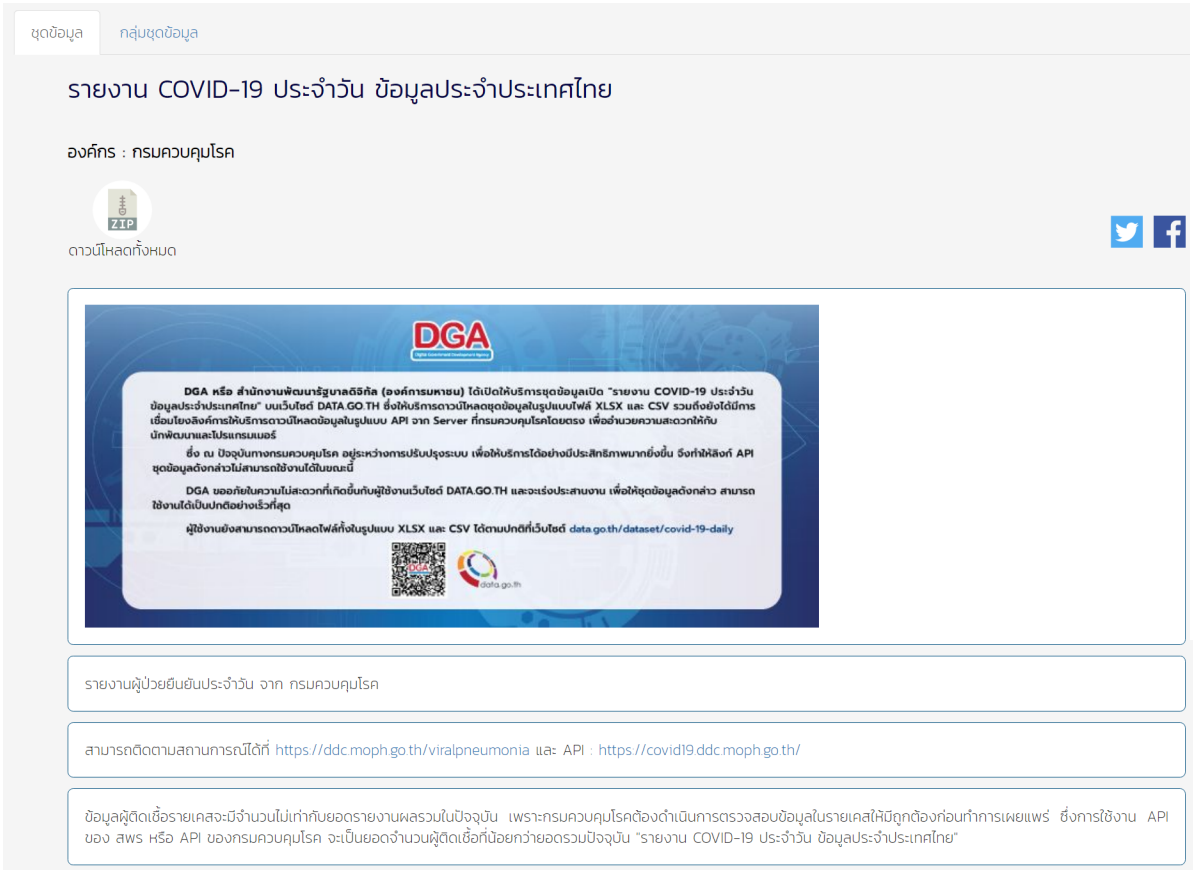
Introduction

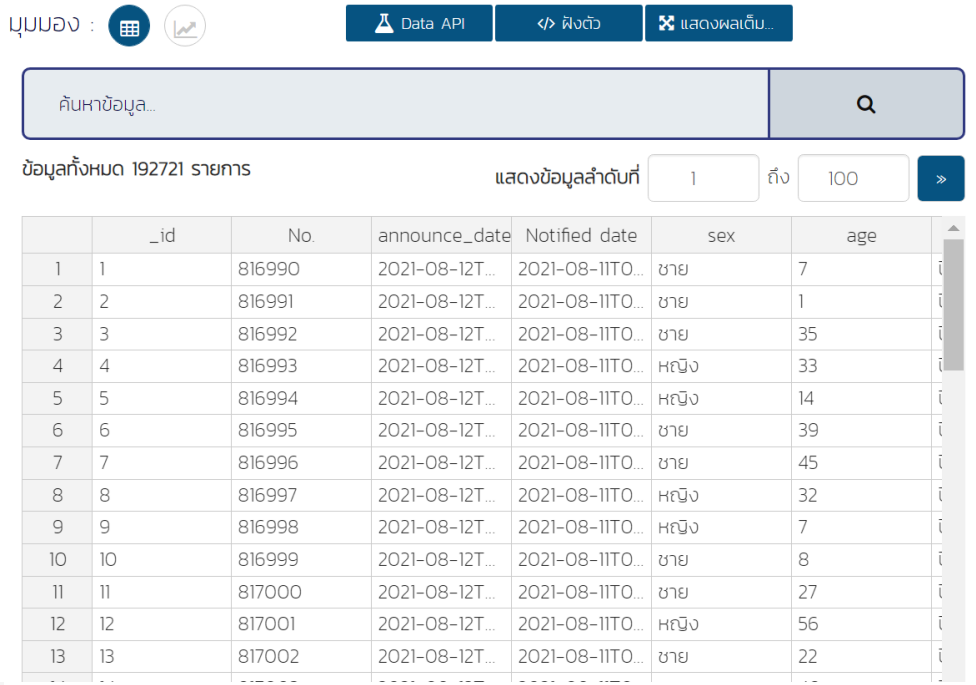
- challenger, grandmaster, master 100,800 game data(korea, 2020)
- There are three large categories
  - team object data
  - participants data
  - gameDuration

# index	# gameCreation game creation code	# gameDuration game time	# gameId game id(match_id)	▵ gameMode game mode
0 … 87.4k	1525b … 1586b	132 … 3.47k	3.19b … 4.26b	CLASSIC 85%, ARAM 11%, Other (3911) 4%
0	1585154663829.0	1323.0	4247263043.0	CLASSIC
1	1585152108929.0	1317.0	4247155821.0	CLASSIC

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

Example: COVID-19 case in Thailand, case report





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

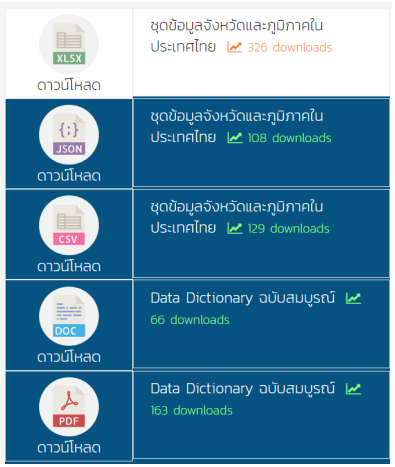
ชุดข้อมูล | กลุ่มชุดข้อมูล

ชุดข้อมูลจังหวัดและภูมิภาคในประเทศไทย

องค์กร : สำนักงานพัฒนารัฐบาลดิจิทัล (องค์การมหาชน)

ZIP
ดาวน์โหลดทั้งหมด

ชุดข้อมูล "การแบ่งภูมิภาคและการแบ่งพื้นที่รับผิดชอบรายจังหวัดของหน่วยงานของรัฐระดับกรมหรือเทียบเท่าจำนวน 69 หน่วยงาน กว่า 90 แบบ ครบทั้ง 77 จังหวัดในประเทศไทย (รวมกรุงเทพมหานคร)" ข้อมูลชุดนี้รวบรวมและตรวจสอบจากแหล่งอ้างอิงที่น่าเชื่อถือ คือ 1. กฎหมายจัดตั้งหน่วยงาน เว็บไซต์ของหน่วยงาน รายงานประจำปีของหน่วยงาน 2. สพร. โทรศัพท์สอบถามเจ้าหน้าที่ของหน่วยงานเพื่อยืนยันความถูกต้องของข้อมูลอีกครั้งหนึ่ง เพื่อตอบคำถามว่า "ประเทศไทยมีกี่ภาค" อย่างมีหลักเกณฑ์และมาตรฐานที่ชัดเจน



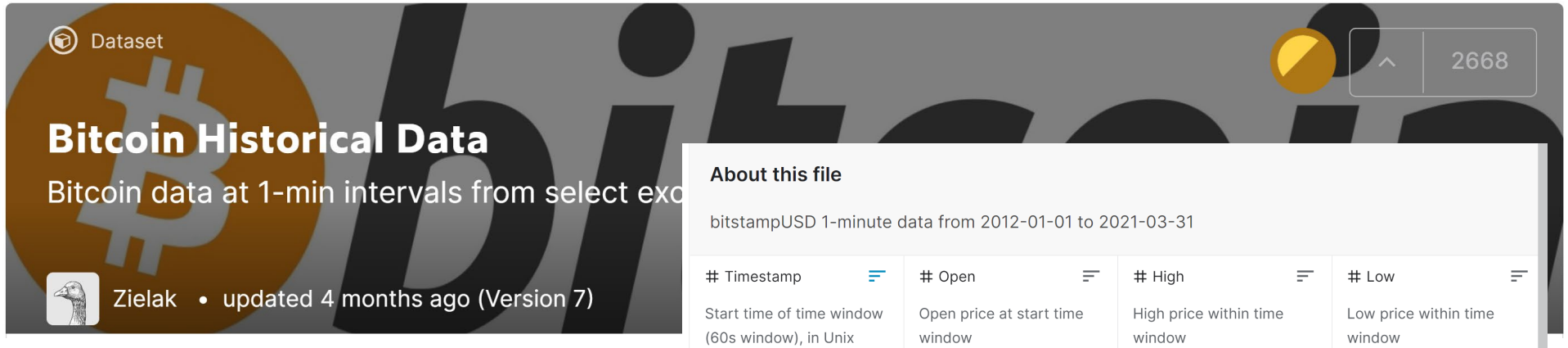
	_id	ProvinceNo	ProvinceMOI_ID	ProvinceNameT	ProvinceNameT	ProvinceBudget
1	1	1	10	กรุงเทพมหาน...	กท	75002
2	2	2	81	จังหวัดกระบี่	กบ	70074
3	3	3	71	จังหวัดกาญจน...	กจ	70041
4	4	4	46	จังหวัดกาฬสินธุ์	กส	70124
5	5	5	62	จังหวัดกำแพงเ...	กพ	70181
6	6	6	40	จังหวัดขอนแก่น	ขก	70122
7	7	7	22	จังหวัดจันทบุรี	จบ	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



About this file

bitstampUSD 1-minute data from 2012-01-01 to 2021-03-31

# Timestamp	# Open	# High	# Low
Start time of time window (60s window), in Unix time	Open price at start time window	High price within time window	Low price within time window
1.33b 1.62b	3.8 61.8k	3.8 61.8k	1.5 61.7k
1521527460	8516.24	8533.92	8509.89
1521527520	8514.57	8514.57	8507.25
1521527580	8513.06	8513.06	8502.4
1521527640	8502.4	8504.64	8497.3

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.

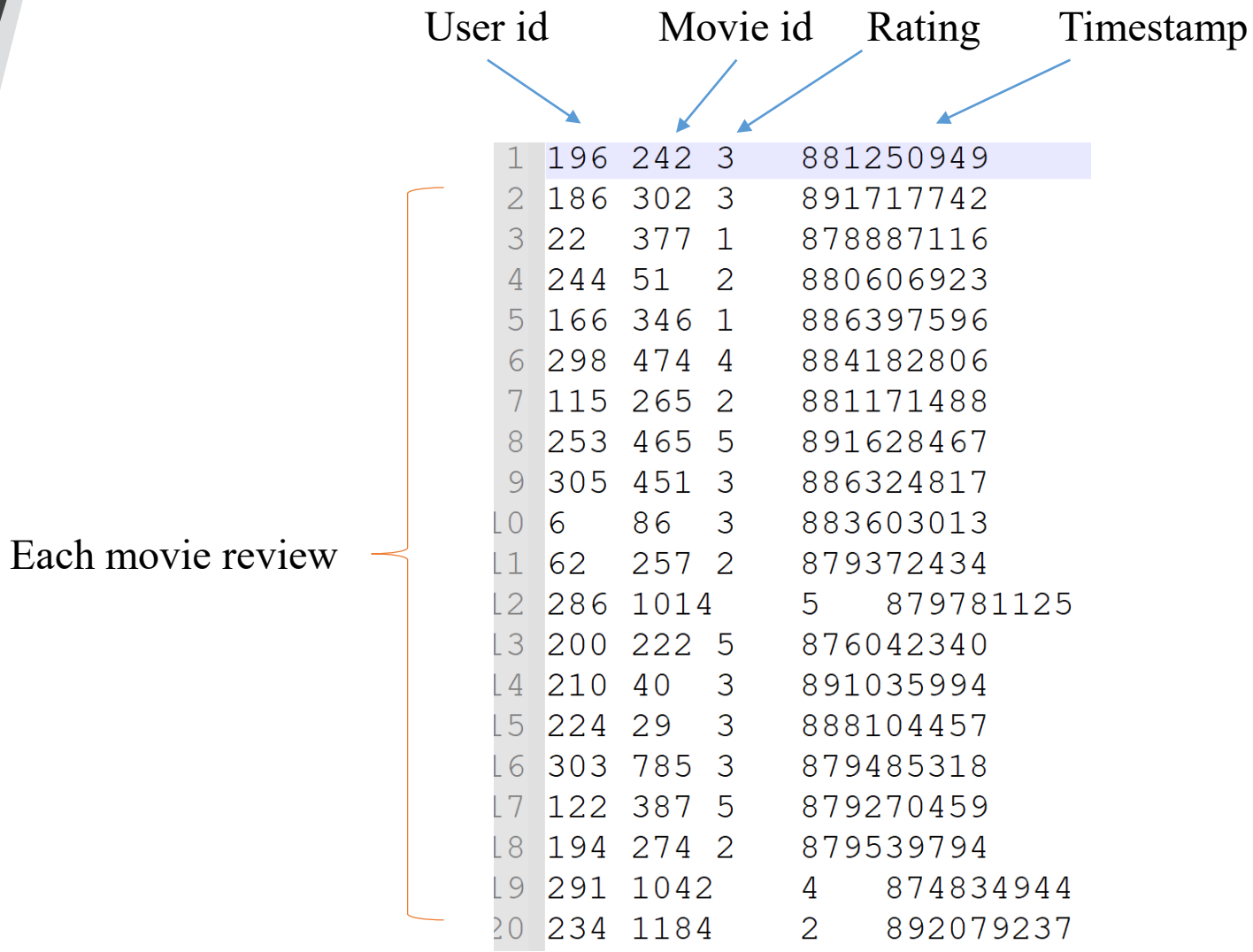
date	meantemp	humidity	wind_speed
Date of format YYYY-MM-DD	Mean temperature averaged out from multiple 3 hour intervals in a day.	Humidity value for the day (units are grams of water vapor per cubic meter volume of air).	Wind speed measured in kmph.
1Jan17 – 24Apr17	11 – 34.5	17.8 – 95.8	1.39 – 19.3
2017-01-01	15.91304347826087	85.8695652173913	2.743478260869565
2017-01-02	18.5	77.22222222222223	2.8944444444444444
2017-01-03	17.11111111111111	81.88888888888889	4.016666666666667
2017-01-04	18.7	70.05	4.545
2017-01-05	18.38888888888889	74.94444444444444	3.3000000000000003

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

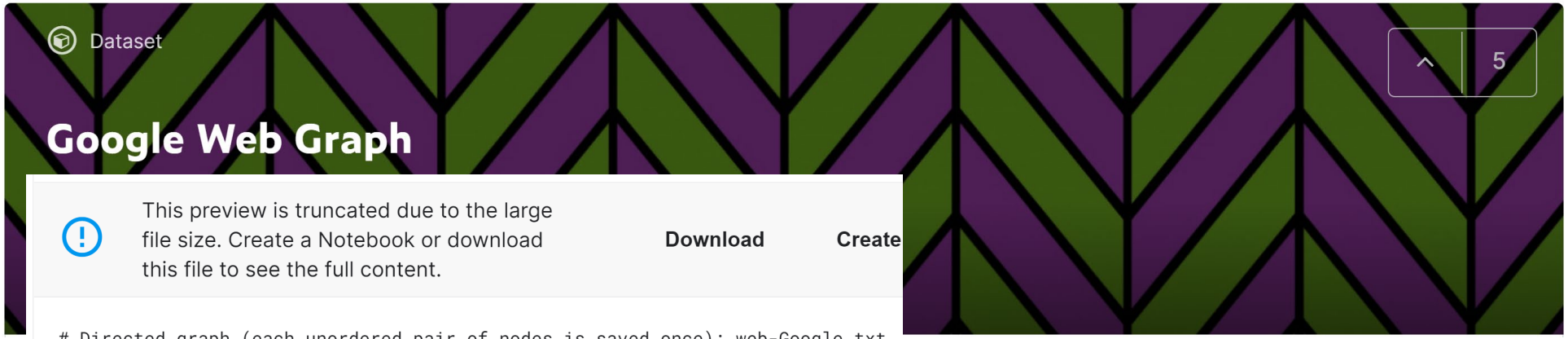
Forms of tabular data
Graph/network data

- 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



Example: Google web hyperlinks



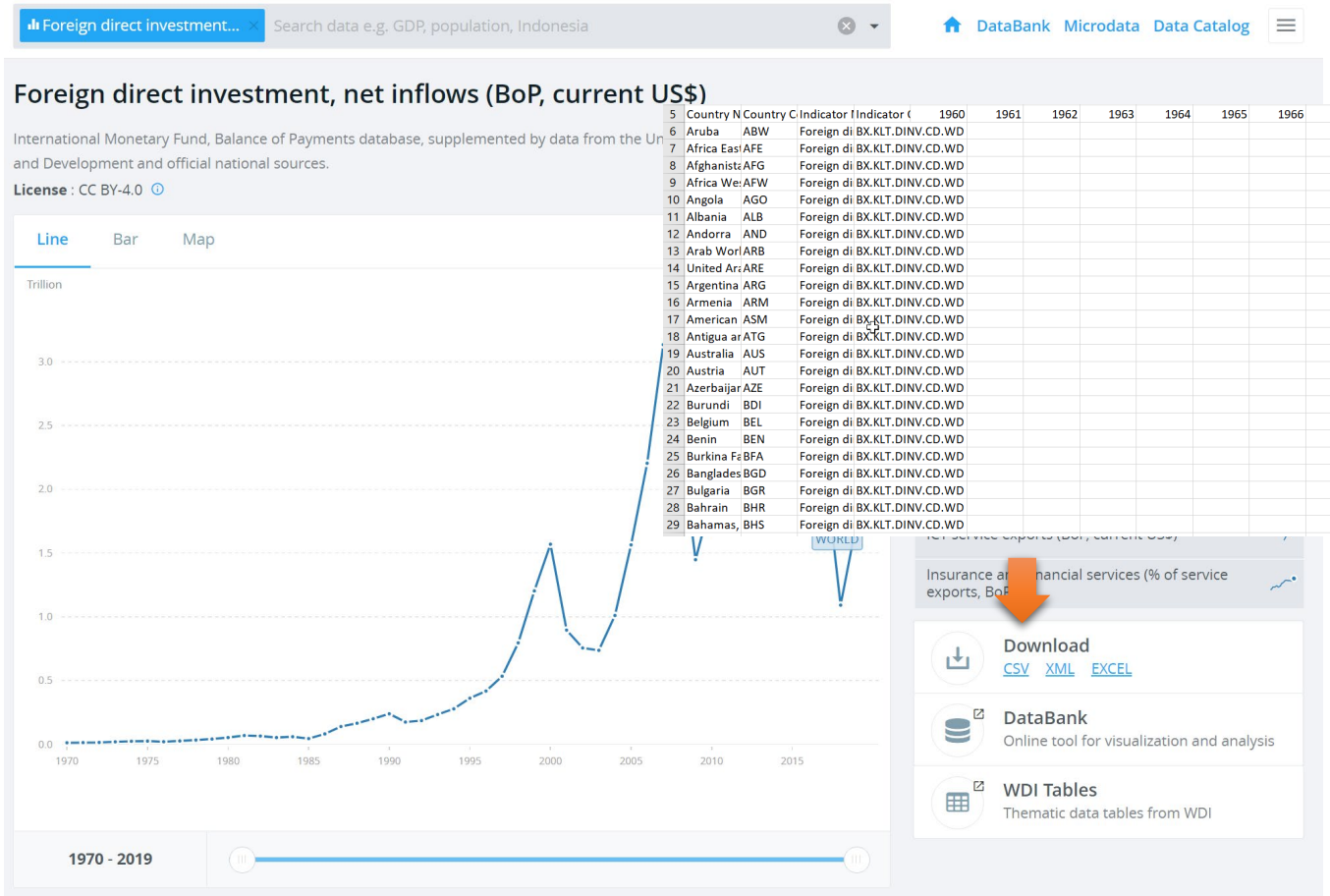

```
# Directed graph (each unordered pair of nodes is saved once): web-Google.txt
# Webgraph from the Google programming contest, 2002
# Nodes: 875713 Edges: 5105039
# FromNodeId    ToNodeId
0       11342
0       824020
0       867923
0       891835
11342   0
11342   27469
11342   38716
11342   309564
11342   322178
11342   387543
11342   427436
11342   538214
11342   638706
```

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

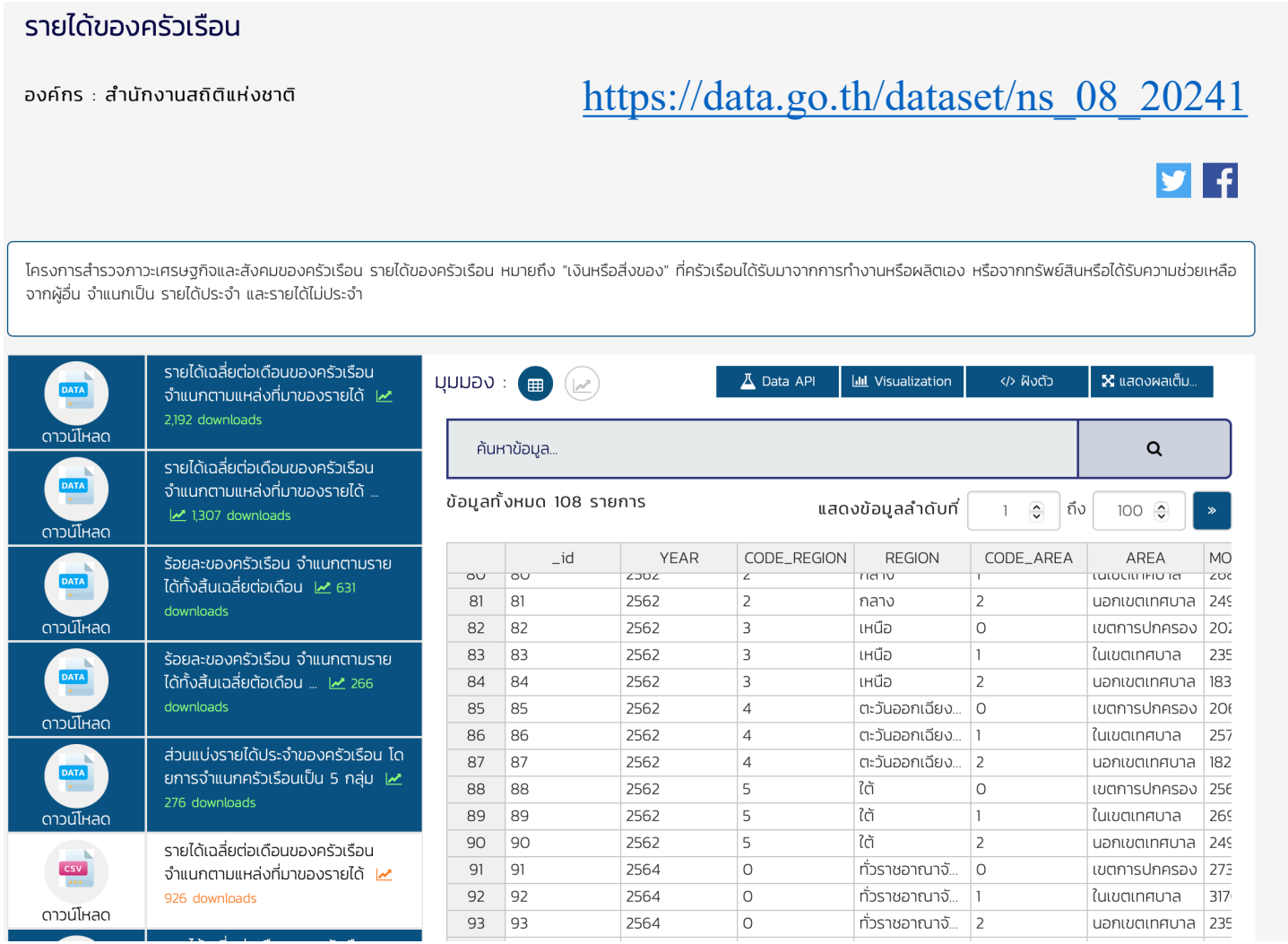
Example: Foreign direct investion, net inflow



	Country N	Country C	Indicator I	Indicator (	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
6	Aruba	ABW	Foreign di	BX.KLT.DINV.CD.WD																				
7	Africa Eas	AFE	Foreign di	BX.KLT.DINV.CD.WD											4.22E+08	3.26E+08	2.19E+08	1.77E+08	8.99E+08	3.17E+08	4.11E+08	67103622	1.68E+08	-1.2E+07
8	Afghanista	AFG	Foreign di	BX.KLT.DINV.CD.WD											230000	450000	150000	270000			40000	-240000		1200000
9	Africa We	AFW	Foreign di	BX.KLT.DINV.CD.WD											4.07E+08	3.89E+08	4.77E+08	6.42E+08	5.18E+08	9.69E+08	6.05E+08	7.79E+08	7.51E+08	7.86E+08
10	Angola	AGO	Foreign di	BX.KLT.DINV.CD.WD											2400000	1860000	2160000	7540000	6690000	50000	-1170000		1870000	2000000
11	Albania	ALB	Foreign di	BX.KLT.DINV.CD.WD																				
12	Andorra	AND	Foreign di	BX.KLT.DINV.CD.WD																				
13	Arab Worl	ARB	Foreign di	BX.KLT.DINV.CD.WD											4.9E+08	47852735	2.68E+08	-6.9E+08	-3.4E+09	1.47E+09	4.07E+08	9.24E+08	6.39E+08	-1.4E+08
14	United Ara	ARE	Foreign di	BX.KLT.DINV.CD.WD											7780000	6940000	10510000	3410000	1630000	15950000	49430000	1.12E+08	39120000	64580000
15	Argentina	ARG	Foreign di	BX.KLT.DINV.CD.WD											89770000	1.26E+08	71720000	1E+08	17920000	55590000	2.45E+08	1.44E+08	2.5E+08	2.06E+08
16	Armenia	ARM	Foreign di	BX.KLT.DINV.CD.WD																				
17	American	ASM	Foreign di	BX.KLT.DINV.CD.WD																				
18	Antigua ar	ATG	Foreign di	BX.KLT.DINV.CD.WD																		2200000	-6900000	8500000
19	Australia	AUS	Foreign di	BX.KLT.DINV.CD.WD											8.93E+08	1.16E+09	1.05E+09	1.35E+08	1.33E+09	4.49E+08	1.05E+09	1.14E+09	1.68E+09	1.49E+09
20	Austria	AUT	Foreign di	BX.KLT.DINV.CD.WD											1.13E+08	1.43E+08	1.45E+08	1.61E+08	1.77E+08	79110000	91230000	96810000	1.42E+08	1.87E+08
21	Azerbaijar	AZE	Foreign di	BX.KLT.DINV.CD.WD																				
22	Burundi	BDI	Foreign di	BX.KLT.DINV.CD.WD											100000		410000	440000	220000	320000	80000	-250000	890000	-300000
23	Belgium	BEL	Foreign di	BX.KLT.DINV.CD.WD											3.14E+08	4.28E+08	3.86E+08	7.19E+08	1.14E+09	9.55E+08	8.72E+08	1.27E+09	1.44E+09	1.13E+09
24	Benin	BEN	Foreign di	BX.KLT.DINV.CD.WD											6700000	2800000	4800000	3500000	-2343120	1889760	2452391	3126022	753359.6	3568043
25	Burkina Fa	BFA	Foreign di	BX.KLT.DINV.CD.WD											400000	1000000	-700000	4400000	2670000	300000	2050000	4970000	1160000	1490000
26	Banglades	BGD	Foreign di	BX.KLT.DINV.CD.WD													90000	2340000	2200000	1543333	5420000	6980000	7700000	-8010000
27	Bulgaria	BGR	Foreign di	BX.KLT.DINV.CD.WD																				
28	Bahrain	BHR	Foreign di	BX.KLT.DINV.CD.WD														3400000	110000	1320000	3210000	650000	22970508	1.45E+08
29	Bahamas,	BHS	Foreign di	BX.KLT.DINV.CD.WD											80320000	54660000	90430000	1.02E+08	1.43E+08	17580000	14600000	31400000	-1100000	9600000

https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD

Example: Income/expense by family



รายได้ของครัวเรือน

องค์กร : สำนักงานสถิติแห่งชาติ

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

โครงการสำรวจภาวะเศรษฐกิจและสังคมของครัวเรือน รายได้ของครัวเรือน หมายถึง "เงินหรือสิ่งของ" ที่ครัวเรือนได้รับมาจากการทำงานหรือผลิตเอง หรือจากทรัพย์สินหรือได้รับความช่วยเหลือจากผู้อื่น จำแนกเป็น รายได้ประจำ และรายได้ไม่ประจำ

ดาวน์โหลด | รายได้เฉลี่ยต่อเดือนของครัวเรือน จำแนกตามแหล่งที่มาของรายได้ 2,192 downloads

ดาวน์โหลด | รายได้เฉลี่ยต่อเดือนของครัวเรือน จำแนกตามแหล่งที่มาของรายได้ ... 1,307 downloads

ดาวน์โหลด | ร้อยละของครัวเรือน จำแนกตามรายได้ทั้งสิ้นเฉลี่ยต่อเดือน 631 downloads

ดาวน์โหลด | ร้อยละของครัวเรือน จำแนกตามรายได้ทั้งสิ้นเฉลี่ยต้อเดือน ... 266 downloads

ดาวน์โหลด | ส่วนแบ่งรายได้ประจำของครัวเรือน โดยการจำแนกครัวเรือนเป็น 5 กลุ่ม 276 downloads

ดาวน์โหลด | รายได้เฉลี่ยต่อเดือนของครัวเรือน จำแนกตามแหล่งที่มาของรายได้ 926 downloads

มุมมอง : Data API | Visualization | ฝังตัว | แสดงผลเต็ม...

ค้นหาข้อมูล...

ข้อมูลทั้งหมด 108 รายการ

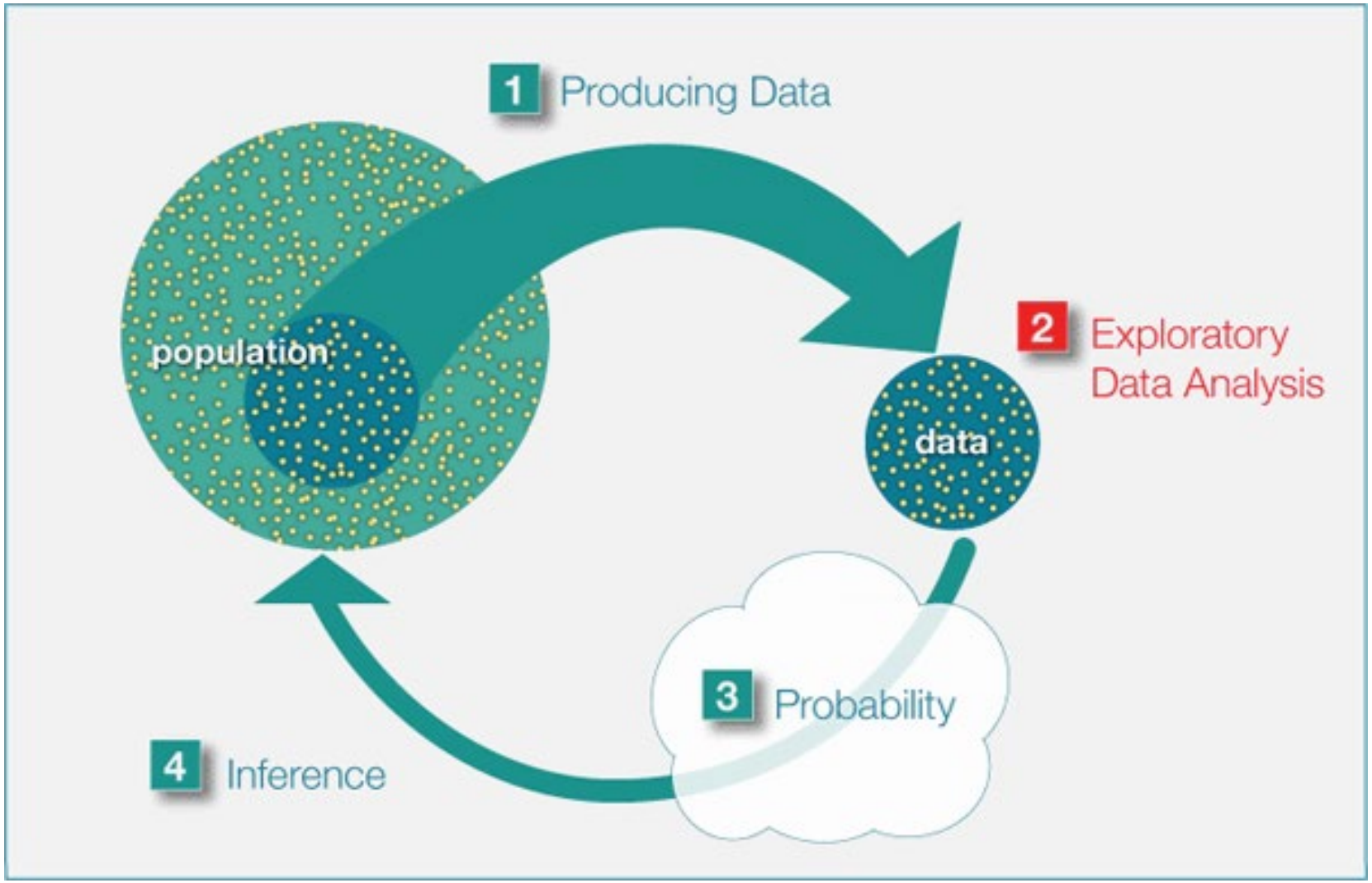
แสดงข้อมูลลำดับที่ 1 ถึง 100

	_id	YEAR	CODE_REGION	REGION	CODE_AREA	AREA	MO
81	81	2562	2	กลาง	2	นอกเขตเทศบาล	249
82	82	2562	3	เหนือ	0	เขตการปกครอง	207
83	83	2562	3	เหนือ	1	ในเขตเทศบาล	235
84	84	2562	3	เหนือ	2	นอกเขตเทศบาล	183
85	85	2562	4	ตะวันออกเฉียง...	0	เขตการปกครอง	206
86	86	2562	4	ตะวันออกเฉียง...	1	ในเขตเทศบาล	257
87	87	2562	4	ตะวันออกเฉียง...	2	นอกเขตเทศบาล	182
88	88	2562	5	ใต้	0	เขตการปกครอง	256
89	89	2562	5	ใต้	1	ในเขตเทศบาล	269
90	90	2562	5	ใต้	2	นอกเขตเทศบาล	249
91	91	2564	0	ทั่วราชอาณาจั...	0	เขตการปกครอง	273
92	92	2564	0	ทั่วราชอาณาจั...	1	ในเขตเทศบาล	317
93	93	2564	0	ทั่วราชอาณาจั...	2	นอกเขตเทศบาล	235

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.head()
```

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	Intern
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	

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

```
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

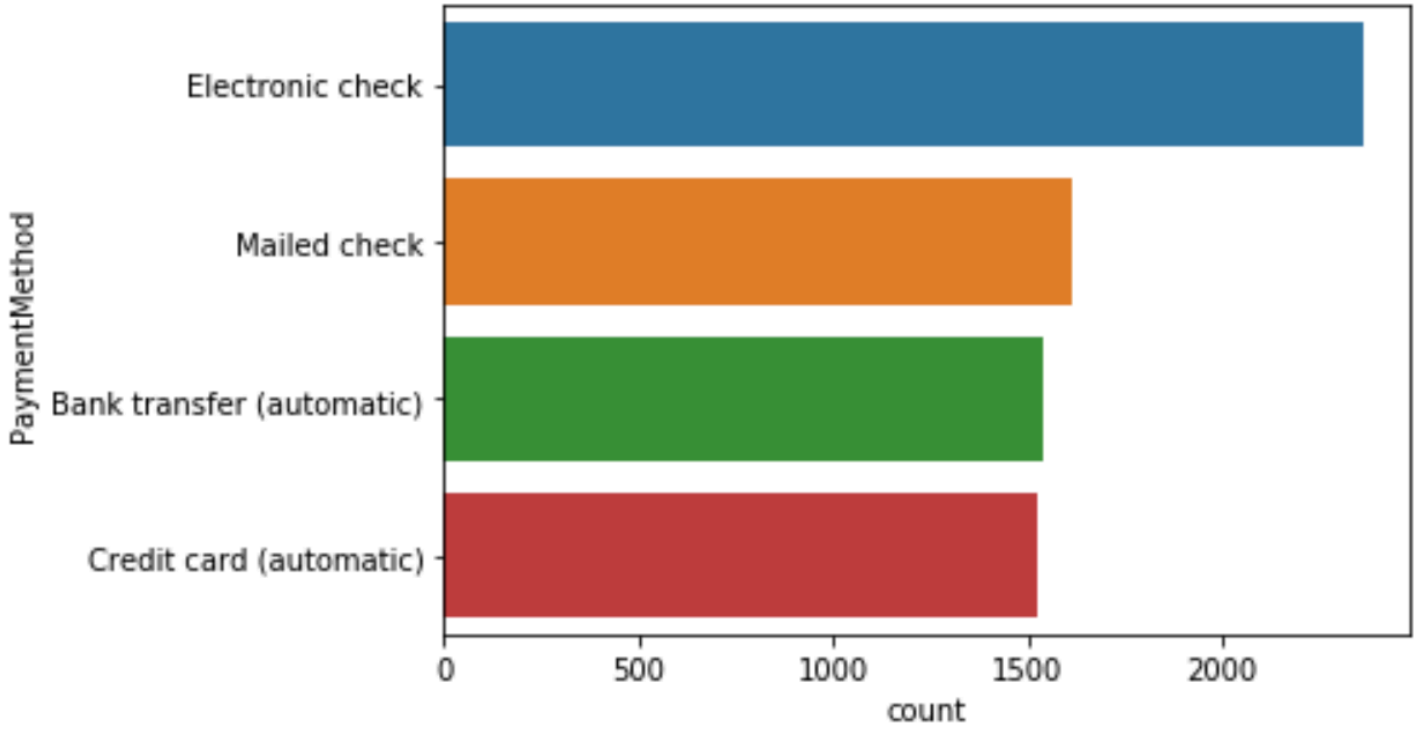
```
pd.crosstab(telcoData['Churn'], columns='Count', normalize='columns')
```

col_0	Count
Churn	
No	0.73463
Yes	0.26537

Countplot

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

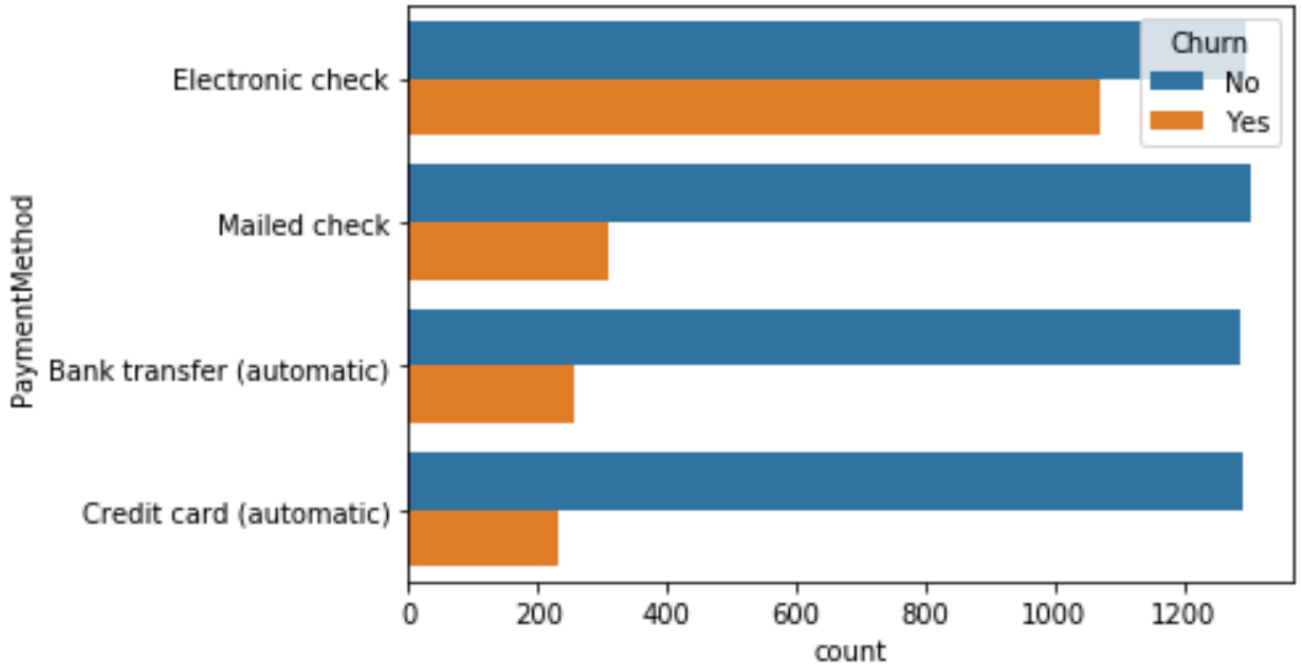
```
<matplotlib.axes._subplots.AxesSubplot at 0x12264ff60>
```



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

BIG DATA EXPERIENCE

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

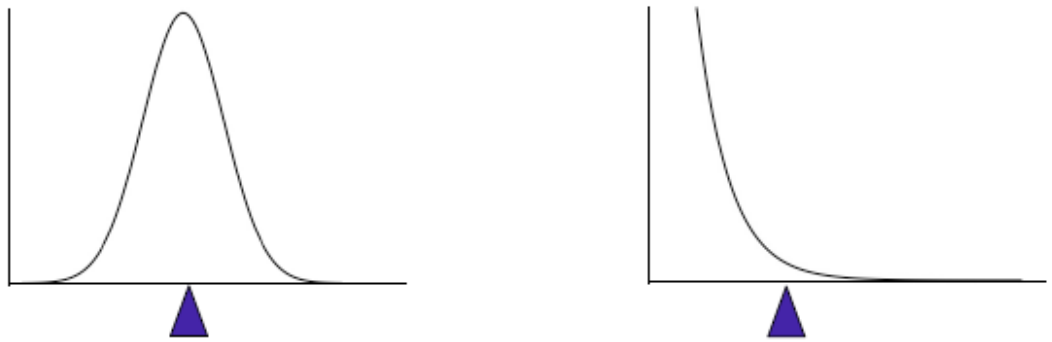
KMUTT

Central tendency
Mean

I. The Mean

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

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



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**

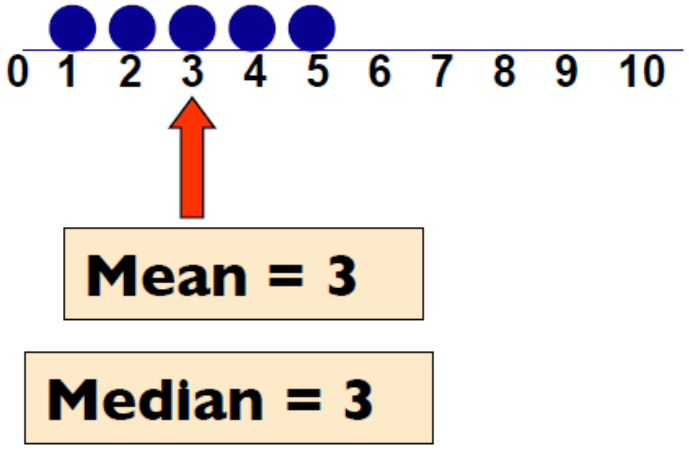
  Some data:
  Age of participants: 17 19 21 22 23 23 23 38

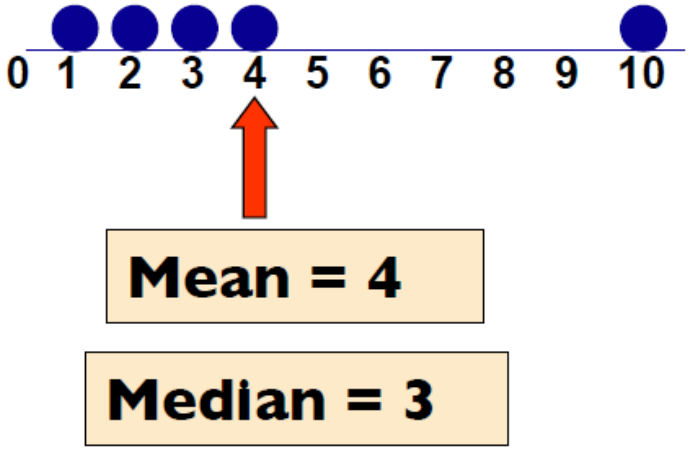
  Median = (22+23)/2 = 22.5

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





Scale: Variance

- Average of squared deviations of values from the mean

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

BIG DATA EXPERIENCE

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

KMUTT

BIG DATA EXPERIENCE

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

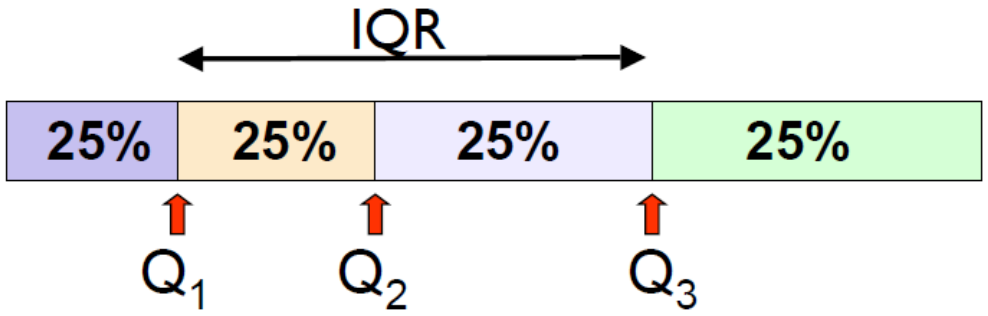
KMUTT

Scale: Standard deviation

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

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!!!

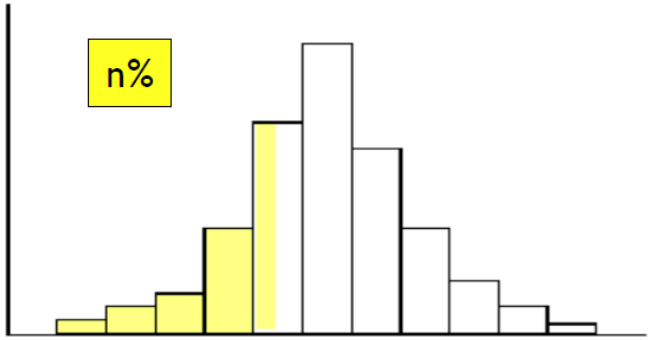
Scale: Quartiles and IQR



- 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 **n^{th} percentile** is a value such that n% of the observations fall at or below or it



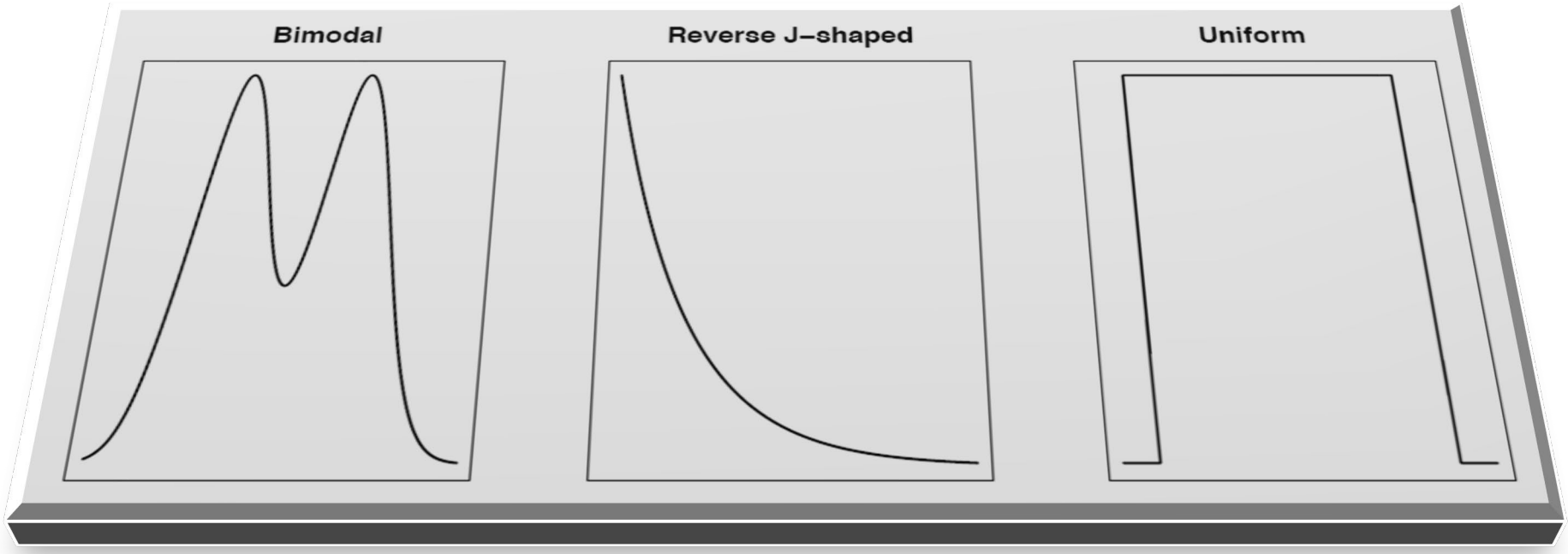
Q_1 = 25th percentile

Median = 50th percentile

Q_2 = 75th percentile

Common distribution shapes

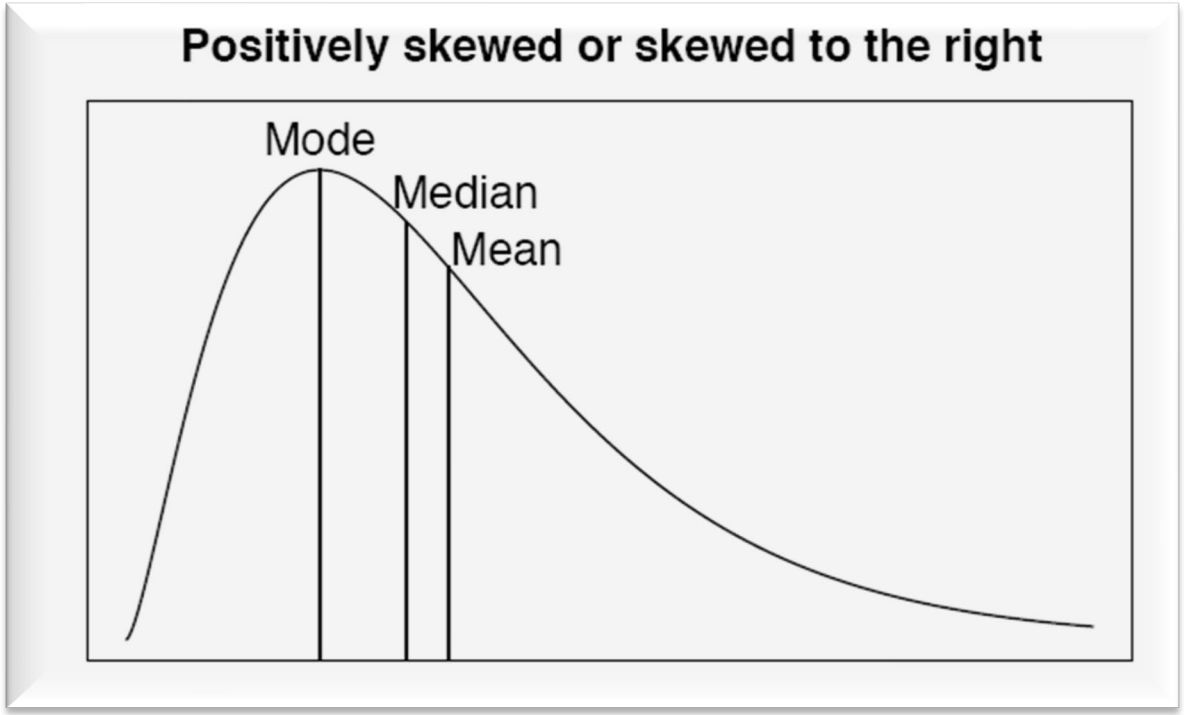
Other distribution shapes



Skewness I

Positively skewed

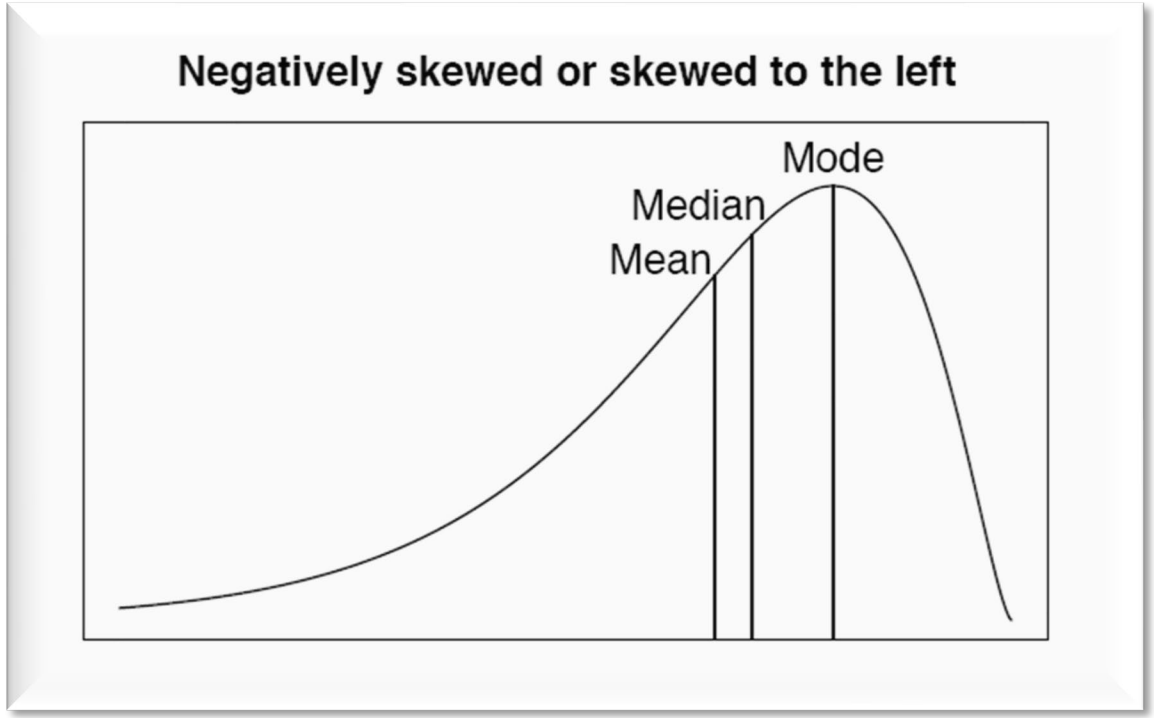
- Longer tail in the high value
- Mean > Median > Mode



Skewness II

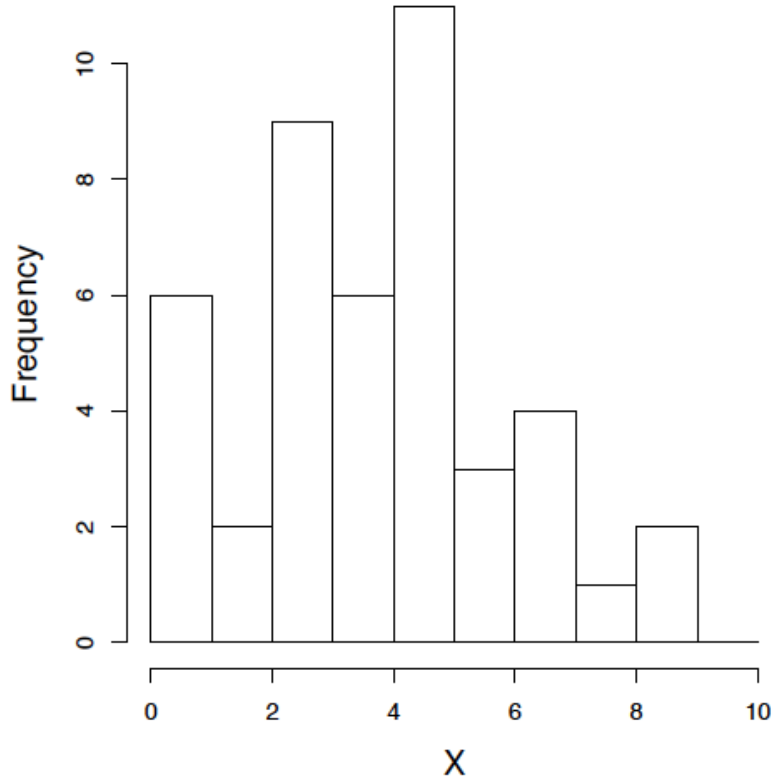
Negatively skewed

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

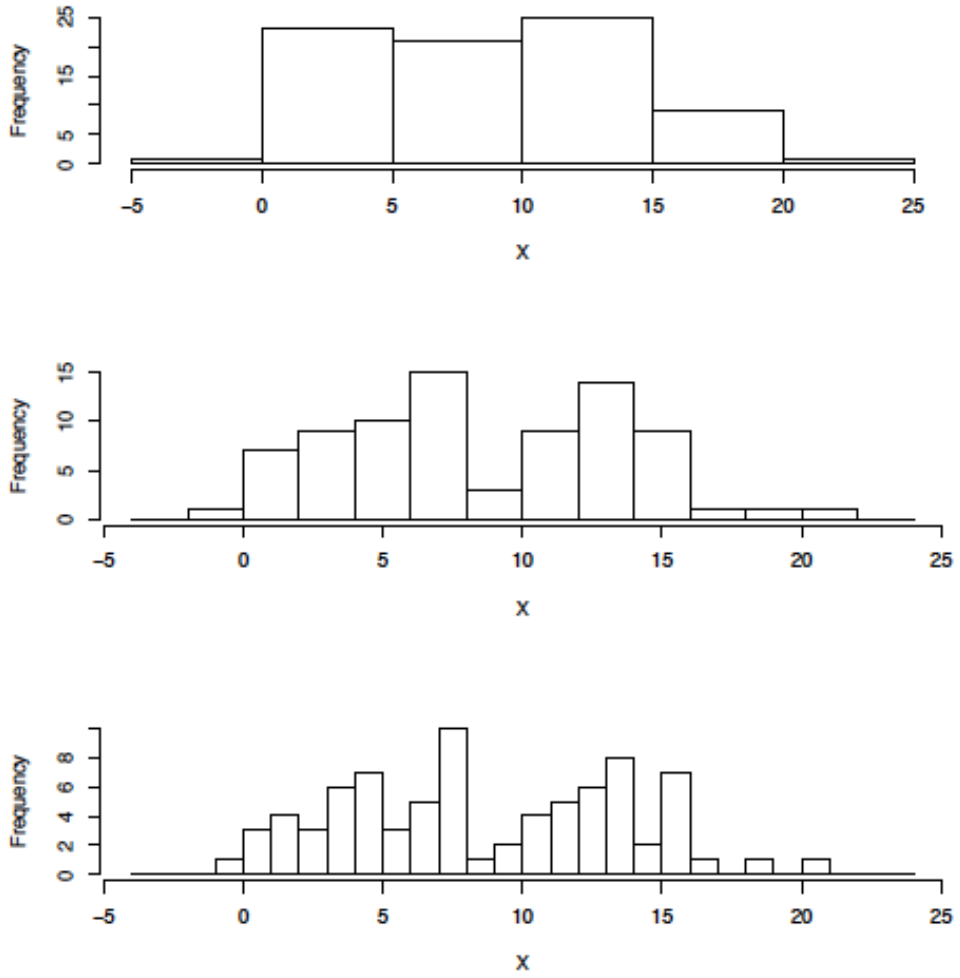


Univariate Graphical EDA
Histogram

- 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.



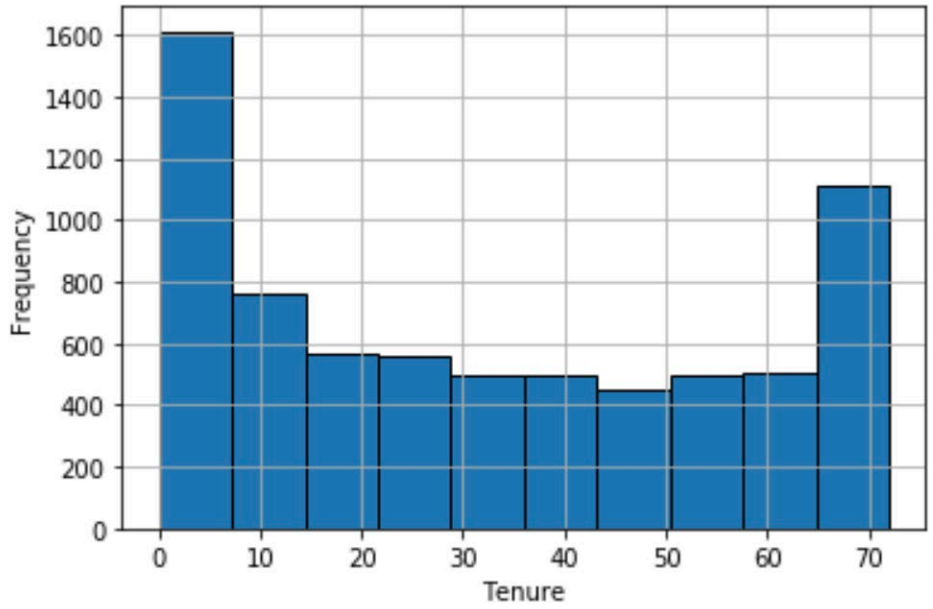
Univariate Graphical EDA
Effects of Histogram Bin



Histogram with Pandas: tenure

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

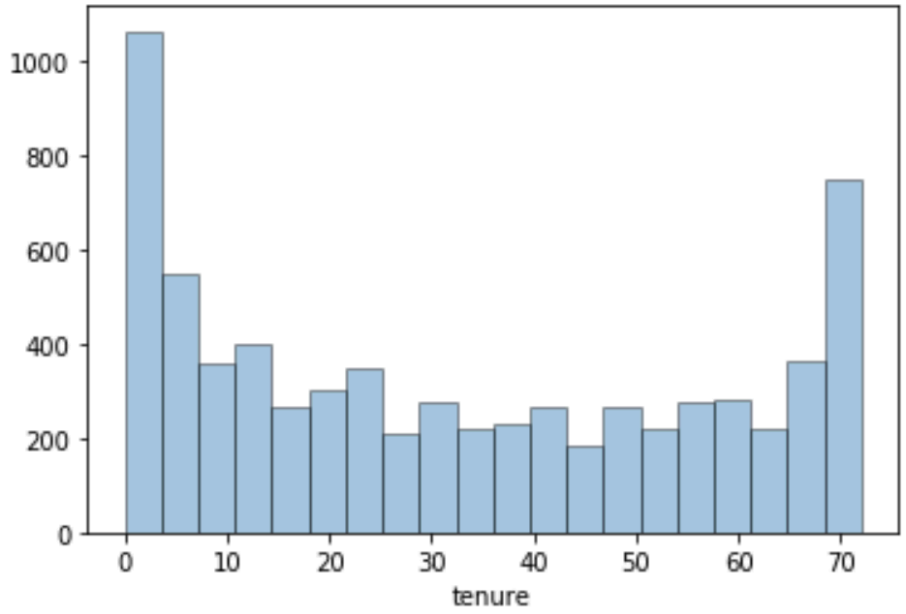
```
Text(0, 0.5, 'Frequency')
```



Histogram with seaborn: tenure

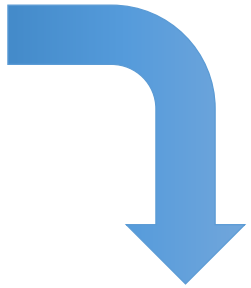
```
sns.distplot(telcoData['tenure'],
             bins=20,
             kde=False,
             hist_kws={'edgecolor':'black'})
```

```
<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

BIG DATA EXPERIENCE

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

KMUTT

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
	Electronic check	0.815562	0.184438
	Mailed check	0.931751	0.068249

BIG DATA EXPERIENCE

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

KMUTT

BIG DATA EXPERIENCE

What to look for?

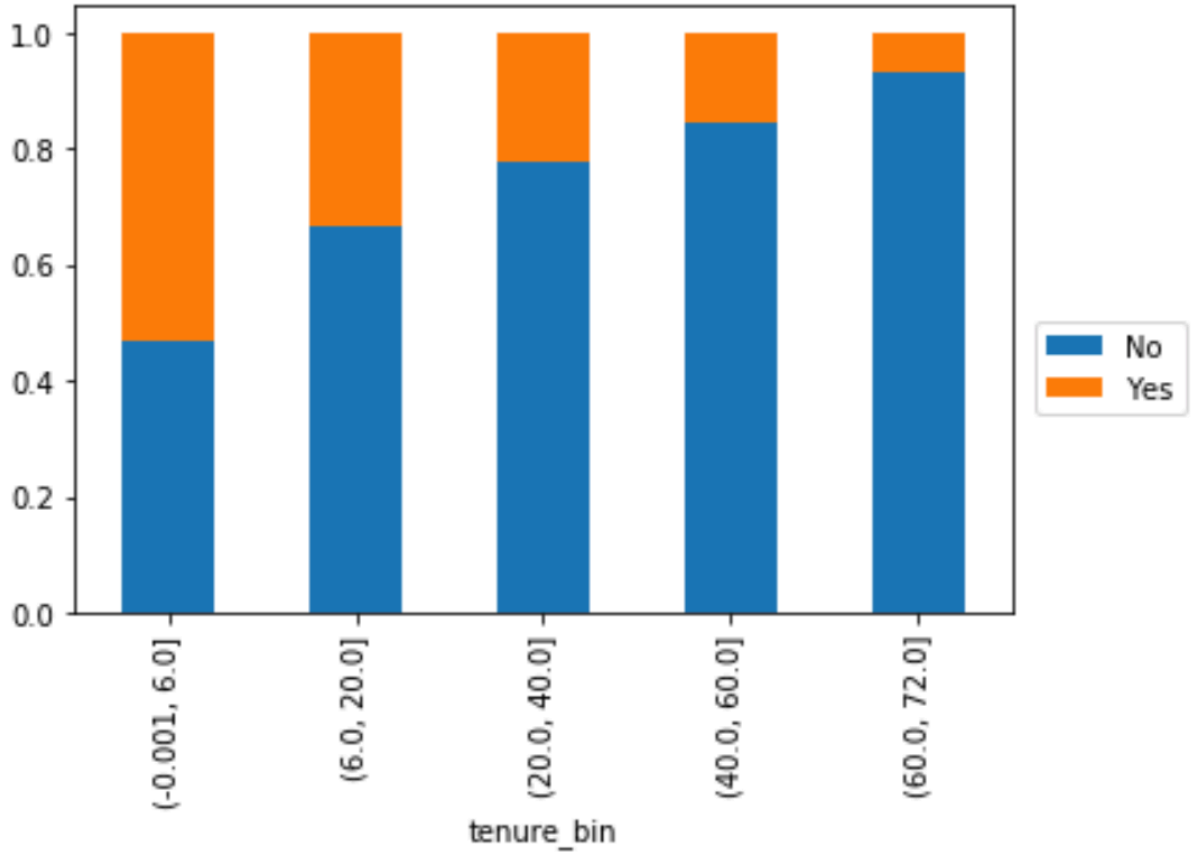
- Relationship to target variable
  - Linear
  - Proportion
  - Specificity
- Dependency
  - Change in distribution when one variable changes

KMUTT

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

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

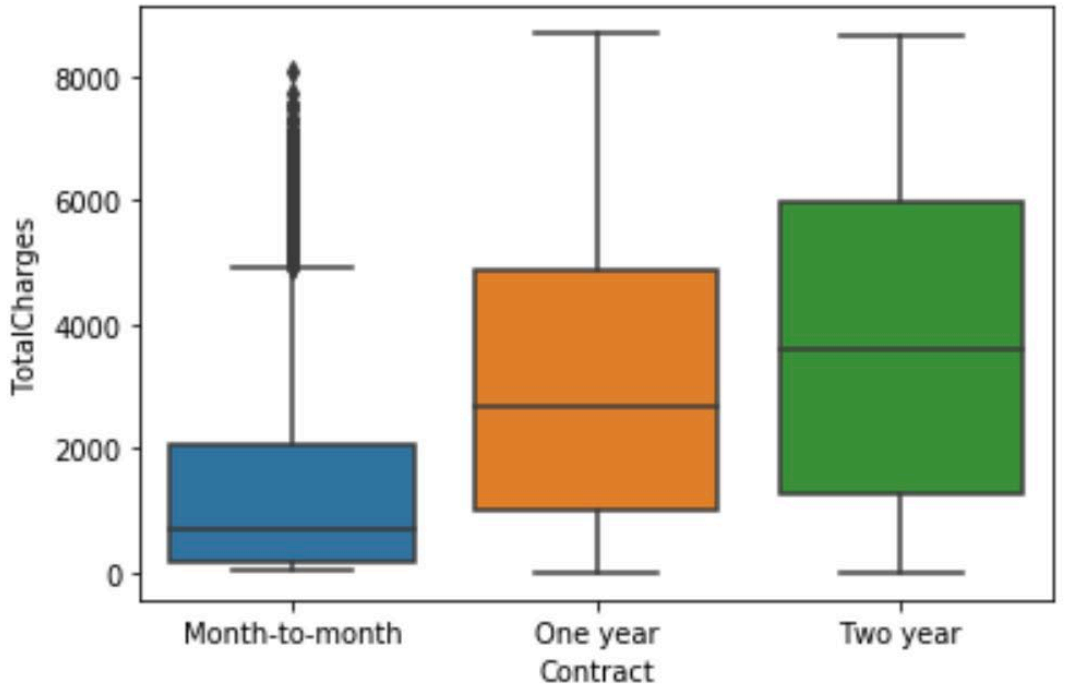
```
<matplotlib.legend.Legend at 0x1a2bc0beb8>
```



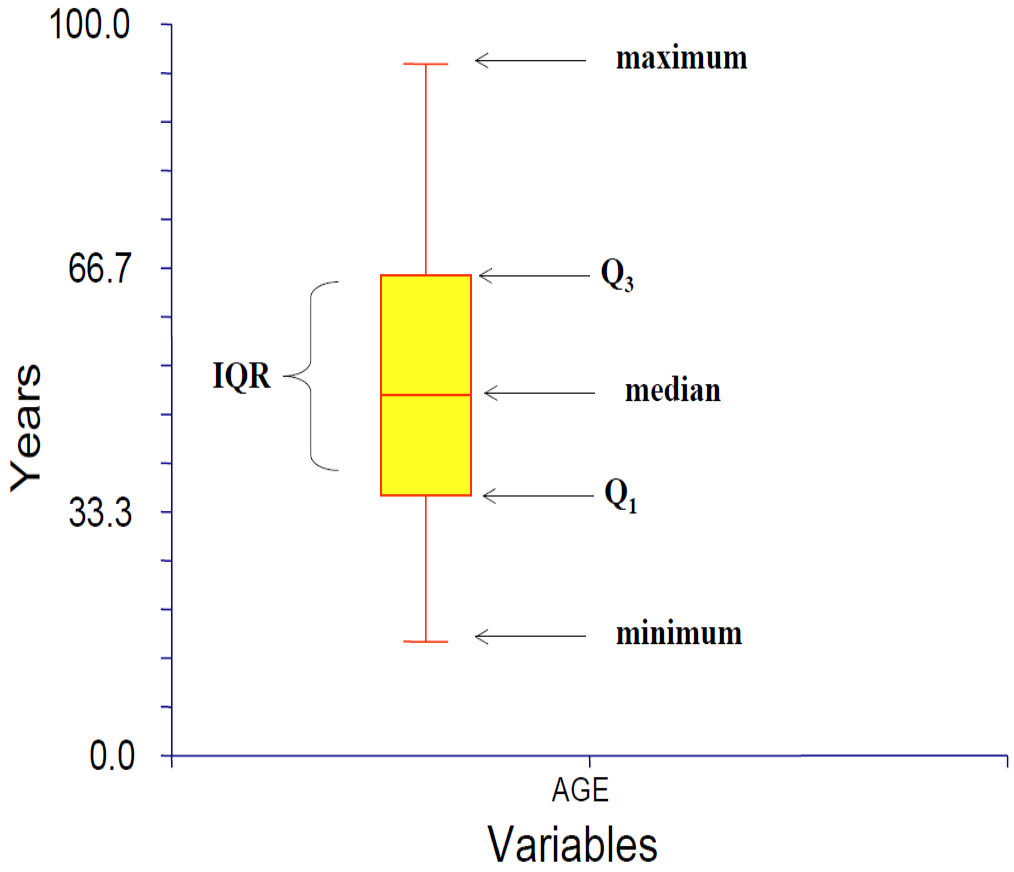
Boxplot

```
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 = Q3 + 1.5*IQR
    - Lower = Q1 – 1.5*IQR

Multivariate analysis
Numerical vs Numerical

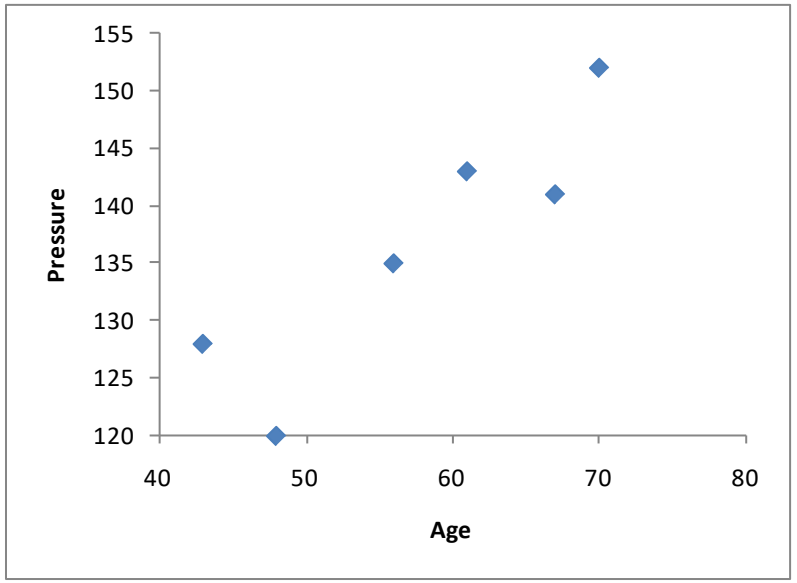
- Graphical
  - Scatter plot

- 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.

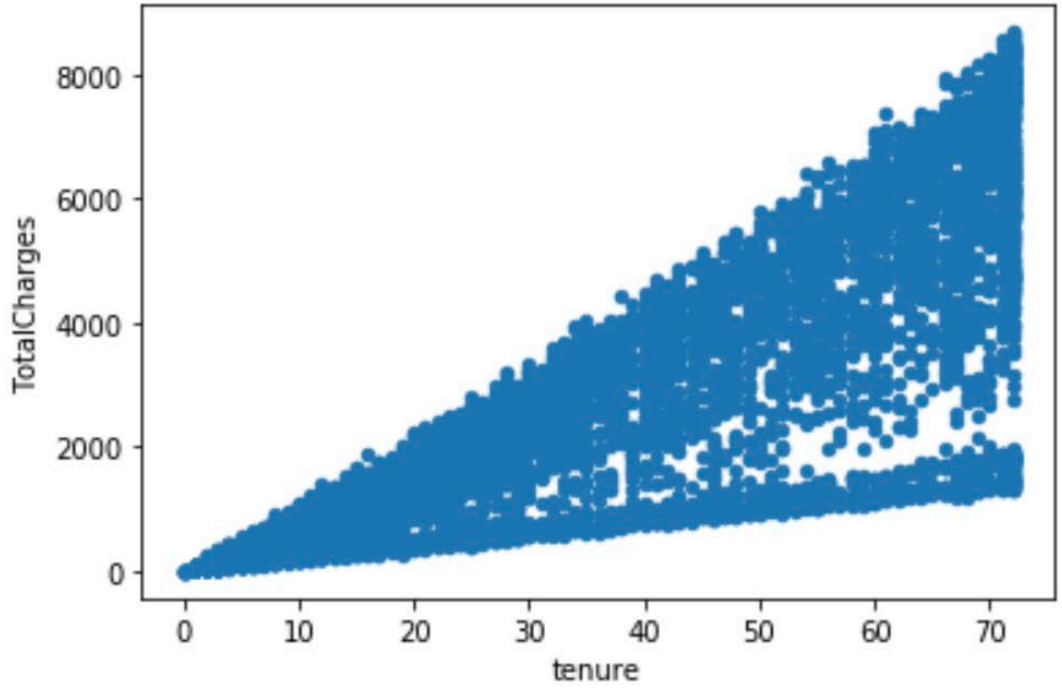
Subject	Age x	Pressure y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152



Scatter plot
Non-aggregated

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

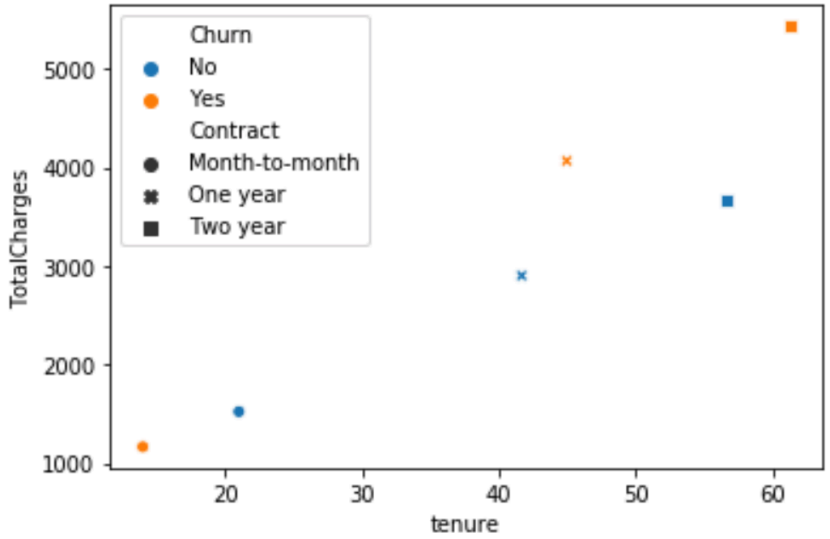
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2dc03c50>
```



Scatter plot Aggregated

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2dfbdf98>
```



BIG DATA EXPERIENCE

Correlation

```
telcoData.corr()
```

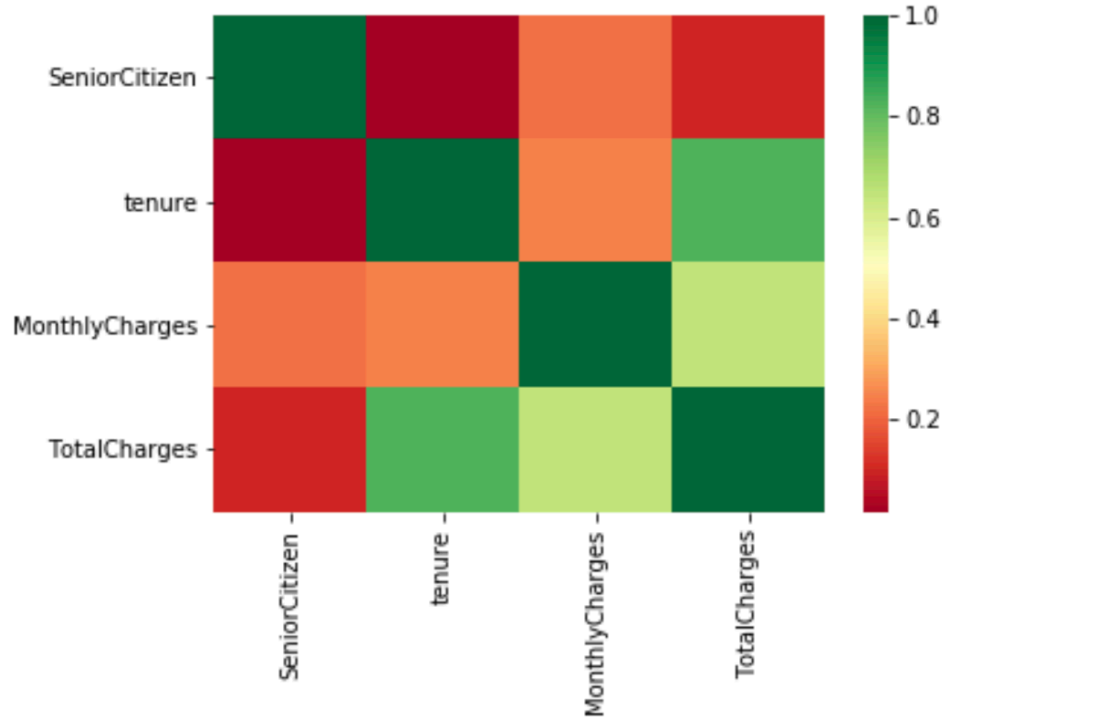
	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
SeniorCitizen	1.000000	0.016567	0.220173	0.103006
tenure	0.016567	1.000000	0.247900	0.826178
MonthlyCharges	0.220173	0.247900	1.000000	0.651174
TotalCharges	0.103006	0.826178	0.651174	1.000000

KMUTT

Heatmap of correlation

```
sns.heatmap(telcoData.corr(), cmap = 'RdYlGn')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2f962048>
```



BIG DATA
EXPERIENCE

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

KMUTT

BIG DATA
EXPERIENCE

Thank you

Question?

KMUTT