Submit work

Data set 1

1. This paper is at https://www.pnas.org/doi/full/10.1073/pnas.0900544106
2. I look for a data set in Refine Bio and found this paper
3. There is 2 dataset in this paper
    1. In the first primate dataset, the people did some dissected postmortem with the DLPFC (a critical region in brain) samples from human, chimpanzee and macaques to extract the RNA. The RNA extract from this process will be store at Affymetrix GC HG-U133 Plus 2.0 arrays.
    2. the second primate dataset was generate by using
4.
    1. There are 20056 features and each feature is each gene in the dorsolateral prefrontal cortex (DLPFC) or caudate nucleus
    2. There are 62 observations.There are 41 observations are from human postmortem brain tissue, 9 observations are from Rhesus macaque dorsolateral prefrontal cortex and 12 from observations are Human caudate nucleus.

Gene	GSM289913	GSM289914	GSM289915	GSM289916	GSM289917	GSM289918	GSM289920
ENSG00000000003	2.4964355759658200	2.5720089233896100	2.752216124187560	1.7266269530827900	2.4363423592233400	1.8104458424908500	2.20656202494049
ENSG00000000005	-0.0904518227541535	0.0044966956285043	0.0451116566501592	0.1161424024032610	0.0780464737681186	0.1349950795729750	0.1270350576825190
ENSG00000000419	7.275908022449870	6.739934696640630	6.980566346497450	7.140912452017360	7.319001023774090	7.330652124647600	7.71479190374985
ENSG00000000457	1.2162341817670500	1.208357379761000	1.1083006799362100	1.0800486282497600	1.3948813766814100	1.3488925571248600	1.2700446939900200
ENSG00000000460	0.024772654077506	-0.0526896459770614	-0.0181384362659103	0.1927624667392750	0.3607524445694660	-0.1139528239521950	-0.114102798378492
ENSG00000000938	0.7204037542824350	0.3561115979425830	0.4271512416697410	0.3311767616240610	0.2775365166630870	0.4198244884445980	0.3383805584150450
ENSG00000000971	0.101575767913062	-0.0492519243595438	0.3788718095443790	0.1202307776898110	0.3878959823356030	0.1058867446737490	-0.0408327376854393
ENSG00000001036	1.0974561719342500	1.2257385298856300	1.2683107810852700	1.8322375211590400	1.4469084344915400	1.0842221875046200	1.8891417195320000
ENSG00000001084	1.6404395356798100	1.5311270532554600	1.4122131341373200	1.6245099955938300	1.23422115691591	1.7603358963278200	1.7414653241041500
ENSG00000001167	0.7026203676472530	0.542487038433695	0.3857403749315420	0.5472878068932620	0.4885949786077220	0.4852147778462560	0.3862691275512950
ENSG00000001460	0.8417333898948300	1.0199829728453800	1.2587328492907700	1.2288702687465500	0.7867109424317590	0.9787364492502010	0.8323919244446680
ENSG00000001461	2.591172039902540	1.988920765723310	1.8856569362878000	2.714670313203130	2.166684095850100	2.82408757499152	2.8099005915723000
ENSG00000001497	1.8658557404194300	1.7104013018372500	2.0234847311417000	1.659661140229710	2.6238299991046700	1.6739777961158600	1.561031845079640
ENSG00000001561	2.182118847246440	2.635432873522360	2.698867896886130	2.717377311885390	2.4055576969606200	3.146197822417190	2.814896783856740
ENSG00000001617	0.8816471163482500	1.1510999261253600	1.0545958076105600	1.045983967542630	1.079275606405040	0.917937222097732	0.7762863553264830
ENSG00000001626	0.0613018843263095	-0.0002246436742277	0.0402396888757831	-0.0523514021780629	-0.112773881168475	0.0022803012838173	0.0265069606257195
ENSG00000001629	7.378863006199920	8.019308509865830	7.697392558135940	7.01238971238404	7.424766797741540	7.954918558535880	6.208221854334870
ENSG00000001631	1.3177659048243000	1.3429930965099500	1.2231217893150000	1.1952544636145700	1.1462565242020900	1.0178392157888700	1.3150238976519800

5. This table contain 20057 rows and 63 columns. Each row is represent for a gene in the brain and each column is represent for collected sample (human postmortem brain tissue, Rhesus macaque dorsolateral prefrontal cortex, caudate nucleus)

Data set 2

1. This paper is at https://ieeexplore.ieee.org/document/9873784
2. I searched for it in google. I look at the method and see they mention the dataset is in Kaggle website, then I went to Kaggle and download the dataset.
3. For this dataset, they collect the information of each house (price, area, number of bedrooms, number of bathroom, etc)
4. a. There are 13 features and they are price, area, bedrooms, bathrooms, stories, main road, guest room, basement, hot water heating, air conditioning, parking, pref area, furnishing status

   b. There are 545 observations and each observation is each house's information collecting session
5. The housing price table contain 13 columns and 546 rows. Each column is each property of a house and each row is each house that the people collected

price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterhea	airconditioni	parking	prefarea	furnishingstatus
13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished
10850000	7500	3	3	1	yes	no	yes	no	yes	2	yes	semi-furnished
10150000	8580	4	3	4	yes	no	no	no	yes	2	yes	semi-furnished
10150000	16200	5	3	2	yes	no	no	no	no	0	no	unfurnished
9870000	8100	4	1	2	yes	yes	yes	no	yes	2	yes	furnished
9800000	5750	3	2	4	yes	yes	no	no	yes	1	yes	unfurnished
9800000	13200	3	1	2	yes	no	yes	no	yes	2	yes	furnished
9681000	6000	4	3	2	yes	yes	yes	yes	no	2	no	semi-furnished
9310000	6550	4	2	2	yes	no	no	no	yes	1	yes	semi-furnished
9240000	3500	4	2	2	yes	no	no	yes	no	2	no	furnished
9240000	7800	3	2	2	yes	no	no	no	no	0	yes	semi-furnished
9100000	6000	4	1	2	yes	no	yes	no	no	2	no	semi-furnished
9100000	6600	4	2	2	yes	yes	yes	no	yes	1	yes	unfurnished
8960000	8500	3	2	4	yes	no	no	no	yes	2	no	furnished
8890000	4600	3	2	2	yes	yes	no	no	yes	2	no	furnished
8855000	6420	3	2	2	yes	no	no	no	yes	1	yes	semi-furnished
8750000	4320	3	1	2	yes	no	yes	yes	no	2	no	semi-furnished
8680000	7155	3	2	1	yes	yes	yes	no	yes	2	no	unfurnished
8645000	8050	3	1	1	yes	yes	yes	no	yes	1	no	furnished
8645000	4560	3	2	2	yes	yes	yes	no	yes	1	no	furnished
8575000	8800	3	2	2	yes	no	no	no	yes	2	no	furnished

Data set 3

1. I found this paper at: https://www.nature.com/articles/s41598-025-15656-8
2. I searched for this paper in Google and found the dataset at the reference section
3. For this dataset, the data collected on the health information of 1500 people. Which is includes age, gender, BMI, Smoking, Genetic Risk, Physical Activity, Alcohol Intake, Cancer History, Diagnosis
4.

1. There are 9 features in this dataset. Which is age, gender, BMI, Smoking, Genetic Risk, Physical Activity, Alcohol Intake, Cancer History, Diagnosis
2. There are 1500 observations and each observation is each patient health information collecting session

# Age	# Gender	# BMI	# Smoking	# GeneticRisk	# PhysicalA...	# AlcoholInt...	# CancerHis...	# Diagnosis
58	1	16.085313321370478	0	1	8.146250560259173	4.148219026764642	1	1
71	0	30.828784389850558	0	1	9.361630415509964	3.519683335172577	0	0
48	1	38.78508355516642	0	2	5.1351786674177005	4.728367685254023	0	1
34	0	30.040295503658278	0	0	9.50279223611407	2.0446361787833602	0	0
62	1	35.47972148566976	0	0	5.356889704560298	3.309849197253171	0	1
27	0	37.1051615837102	0	1	3.9419049459990796	2.3242739040663536	0	0
80	1	20.701994298723314	0	0	8.482031285528954	3.1529434865427177	0	0
40	0	20.301120997705667	1	0	4.929827104580891	2.2479945300386865	1	0
58	1	30.274524721274794	0	1	4.719025105645231	0.9431612321203742	1	1
77	0	25.27571174516762	0	1	4.9031248940982834	2.5394870800169045	0	0
38	1	35.99653257166801	0	2	8.467850018697455	2.6429210071872116	0	1
42	0	37.50057808495028	1	2	8.319415416880052	4.053019241300805	0	1
30	1	23.835534482567795	0	1	4.035199924402112	3.677603528967906	1	1

5. This table contain 9 rows and 1500 columns. Each row is represents for patient and each column is represents for a health record session