

به نام خدا

تمرین اول

مبانی داده کاوی

شناخت داده

(انتخاب داده، بررسی نوع و توصیف آماری)

نام و نام خانوادگی: علی رضائی نژاد

شماره دانشجویی: ۹۶۰۱۸۴۱۵۶

مشخصه درس: ۹۱۳۵۱

نام استاد: خانم امینه امینی

انتخاب مجموعه داده

ابتدا از آدرس [UCI Machine Learning Repository: Data Sets](#) یک مجموعه داده (دیتاست)، که برای طبقه‌بندی برچسب داشته باشد و برای جلوگیری از پیچیدگی و نیاز به حذف، ابعاد آن بیش از حد بزرگ نباشد، را انتخاب می‌کنیم.


Default Task - Undo
Classification (41)
Regression (11)
Clustering (12)
Other (6)
Attribute Type
Categorical (6)
Numerical (24)
Mixed (3)
Data Type
Multivariate (33)
Univariate (3)
Sequential (0)
Time-Series (3)
Text (6)
Domain-Theory (2)
Other (0)
Area
Life Sciences (14)
Physical Sciences (1)
CS / Engineering (11)
Social Sciences (3)
Business (3)
Game (1)
Other (7)
Attributes - Undo
Less than 10 (41)
10 to 100 (92)
Greater than 100 (20)
Instances - Undo
Less than 100 (13)
100 to 1000 (41)
Greater than 1000 (54)
Format Type
Matrix (29)
Non-Matrix (12)

با استفاده از سه دسته‌بندی **Default Task**، **Attributes** و **Instances** داده‌هایی برچسب‌دار با کم‌تر از ۱۰ ویژگی (فیلد) و بین ۱۰۰ تا ۱۰۰۰ نمونه را تفکیک کردم که حاصل آن ۴۱ دیتاست شد. (تصویر روبه‌رو)

از میان دیتاست‌های موجود، مجموعه **Seeds** که شامل بررسی ۲۱۰ دانه گندم از سه نوع مختلف (**Rosa**، **Kama**) و **Canadian** (و ابعاد هر دانه (طول و عرض)، مساحت، محیط و طول **groove** دانه‌ها، **Compactness** ([چگالی](#)))، ضریب عدم تقارن (**Asymmetry Coefficient**) آن‌ها بود را انتخاب کردم.

از هر سه نژاد ۷۰ نمونه در دیتاست (۳۳.۳ درصد از تمام نمونه‌ها) گنجانده شده است.

[UCI Machine Learning Repository: seeds Data Set](#)

 seeds	Multivariate	Classification, Clustering	Real	210	7	2012
---	--------------	----------------------------	------	-----	---	------

بررسی صفات

هفت صفت طول، عرض، مساحت، محیط، Compactness (دنسیتی/چگالی)، ضریب عدم تقارن دانه و طول groove دانه‌ها از نوع عددی و Ratio-Scaled هستند، فیلد هشتم شامل کلاس (ویژگی اسمی یا categorical) که بیانگر یکی از سه نژاد بررسی شده است، می‌باشد.

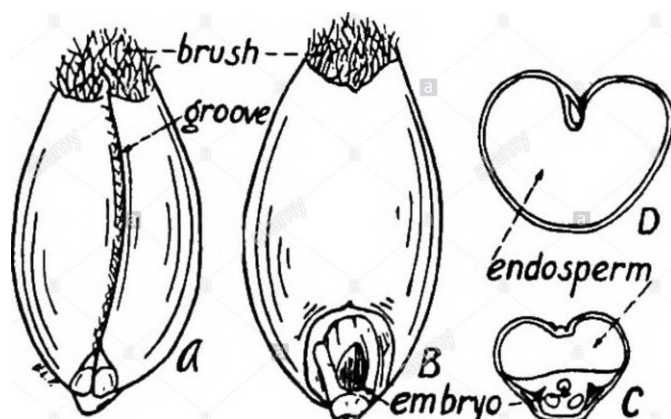
نمونه‌ای از یک رکورد:

15.26 14.84 0.871 5.763 3.312 2.221 5.22 1

برای داده‌ی اسمی به صورت زیر معادل عددی در نظر گرفته شده است. زین پس با عدد مربوط به عنوان نژاد بررسی شده در هر اندازه‌گیری که از پیش مهیا بود ادامه می‌دهیم.

Class1 = Kama **Class2 = Rosa** **Class3 = Canadian**

هر هفت ویژگی عددی، پیوسته بوده و همه ویژگی‌ها جز Compactness و ضریب عدم تقارن در مقیاس میلی‌متر/میلی‌متر مربع و به ترتیب از چپ به راست بیانگر مساحت دانه، محیط دانه، Compactness، طول هسته، عرض هسته، ضریب عدم تقارن (Asymmetry Coefficient) و درازای groove دانه می‌باشند.



منظور از groove

در دانه‌ی گندم:

همانطور که در نمونه رکورد صفحه قبل مشخص است مقادیر دیتاست با فاصله و تب (Tab) از یکدیگر جدا شده‌اند. پیش از ادامه کار مجموعه داده را با استفاده از مایکروسافت اکسل به فرمت جدا شده با کاما (CSV) بدل کردم.

خلاصه‌ای از دیتاست و نمایش ساختار آن:

	A	P	C	L	W	AC	LG	Class
1	15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
3	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
.
.
.
210	12.3	13.34	0.8684	5.243	2.974	5.637	5.063	3

- | | | |
|-------|---------------------------|----------------------|
| 1. A | = Area | مساحت |
| 2. P | = Perimeter | محیط |
| 3. C | = Compactness | دنسیتی/چگالی کل دانه |
| 4. L | = Length of Kernel | طول هسته |
| 5. W | = Width of Kernel | عرض هسته |
| 6. AC | = Asymmetry Coefficient | ضریب عدم تقارن |
| 7. LG | = Length of Kernel Groove | طول گروو دانه |

نمای کامل مجموعه داده در قالب جدول

<i>Instances</i>	<i>Area</i>	<i>Perimeter</i>	<i>Compactness</i>	<i>L</i>	<i>W</i>	<i>AC</i>	<i>LG</i>	<i>C</i>
1)	15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
2)	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
3)	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
4)	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
5)	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1
6)	14.38	14.21	0.8951	5.386	3.312	2.462	4.956	1
7)	14.69	14.49	0.8799	5.563	3.259	3.586	5.219	1
8)	14.11	14.1	0.8911	5.42	3.302	2.7	5	1
9)	16.63	15.46	0.8747	6.053	3.465	2.04	5.877	1
10)	16.44	15.25	0.888	5.884	3.505	1.969	5.533	1
11)	15.26	14.85	0.8696	5.714	3.242	4.543	5.314	1
12)	14.03	14.16	0.8796	5.438	3.201	1.717	5.001	1
13)	13.89	14.02	0.888	5.439	3.199	3.986	4.738	1
14)	13.78	14.06	0.8759	5.479	3.156	3.136	4.872	1
15)	13.74	14.05	0.8744	5.482	3.114	2.932	4.825	1
16)	14.59	14.28	0.8993	5.351	3.333	4.185	4.781	1
17)	13.99	13.83	0.9183	5.119	3.383	5.234	4.781	1
18)	15.69	14.75	0.9058	5.527	3.514	1.599	5.046	1
19)	14.7	14.21	0.9153	5.205	3.466	1.767	4.649	1
20)	12.72	13.57	0.8686	5.226	3.049	4.102	4.914	1
21)	14.16	14.4	0.8584	5.658	3.129	3.072	5.176	1
22)	14.11	14.26	0.8722	5.52	3.168	2.688	5.219	1
23)	15.88	14.9	0.8988	5.618	3.507	0.7651	5.091	1
24)	12.08	13.23	0.8664	5.099	2.936	1.415	4.961	1
25)	15.01	14.76	0.8657	5.789	3.245	1.791	5.001	1
26)	16.19	15.16	0.8849	5.833	3.421	0.903	5.307	1
27)	13.02	13.76	0.8641	5.395	3.026	3.373	4.825	1
28)	12.74	13.67	0.8564	5.395	2.956	2.504	4.869	1
29)	14.11	14.18	0.882	5.541	3.221	2.754	5.038	1
30)	13.45	14.02	0.8604	5.516	3.065	3.531	5.097	1
31)	13.16	13.82	0.8662	5.454	2.975	0.8551	5.056	1
32)	15.49	14.94	0.8724	5.757	3.371	3.412	5.228	1
33)	14.09	14.41	0.8529	5.717	3.186	3.92	5.299	1
34)	13.94	14.17	0.8728	5.585	3.15	2.124	5.012	1

35)	15.05	14.68	0.8779	5.712	3.328	2.129	5.36	1
36)	16.12	15	0.9	5.709	3.485	2.27	5.443	1
37)	16.2	15.27	0.8734	5.826	3.464	2.823	5.527	1
38)	17.08	15.38	0.9079	5.832	3.683	2.956	5.484	1
39)	14.8	14.52	0.8823	5.656	3.288	3.112	5.309	1
40)	14.28	14.17	0.8944	5.397	3.298	6.685	5.001	1
41)	13.54	13.85	0.8871	5.348	3.156	2.587	5.178	1
42)	13.5	13.85	0.8852	5.351	3.158	2.249	5.176	1
43)	13.16	13.55	0.9009	5.138	3.201	2.461	4.783	1
44)	15.5	14.86	0.882	5.877	3.396	4.711	5.528	1
45)	15.11	14.54	0.8986	5.579	3.462	3.128	5.18	1
46)	13.8	14.04	0.8794	5.376	3.155	1.56	4.961	1
47)	15.36	14.76	0.8861	5.701	3.393	1.367	5.132	1
48)	14.99	14.56	0.8883	5.57	3.377	2.958	5.175	1
49)	14.79	14.52	0.8819	5.545	3.291	2.704	5.111	1
50)	14.86	14.67	0.8676	5.678	3.258	2.129	5.351	1
51)	14.43	14.4	0.8751	5.585	3.272	3.975	5.144	1
52)	15.78	14.91	0.8923	5.674	3.434	5.593	5.136	1
53)	14.49	14.61	0.8538	5.715	3.113	4.116	5.396	1
54)	14.33	14.28	0.8831	5.504	3.199	3.328	5.224	1
55)	14.52	14.6	0.8557	5.741	3.113	1.481	5.487	1
56)	15.03	14.77	0.8658	5.702	3.212	1.933	5.439	1
57)	14.46	14.35	0.8818	5.388	3.377	2.802	5.044	1
58)	14.92	14.43	0.9006	5.384	3.412	1.142	5.088	1
59)	15.38	14.77	0.8857	5.662	3.419	1.999	5.222	1
60)	12.11	13.47	0.8392	5.159	3.032	1.502	4.519	1
61)	11.42	12.86	0.8683	5.008	2.85	2.7	4.607	1
62)	11.23	12.63	0.884	4.902	2.879	2.269	4.703	1
63)	12.36	13.19	0.8923	5.076	3.042	3.22	4.605	1
64)	13.22	13.84	0.868	5.395	3.07	4.157	5.088	1
65)	12.78	13.57	0.8716	5.262	3.026	1.176	4.782	1
66)	12.88	13.5	0.8879	5.139	3.119	2.352	4.607	1
67)	14.34	14.37	0.8726	5.63	3.19	1.313	5.15	1
68)	14.01	14.29	0.8625	5.609	3.158	2.217	5.132	1
69)	14.37	14.39	0.8726	5.569	3.153	1.464	5.3	1
70)	12.73	13.75	0.8458	5.412	2.882	3.533	5.067	1
71)	17.63	15.98	0.8673	6.191	3.561	4.076	6.06	2

72)	16.84	15.67	0.8623	5.998	3.484	4.675	5.877	2
73)	17.26	15.73	0.8763	5.978	3.594	4.539	5.791	2
74)	19.11	16.26	0.9081	6.154	3.93	2.936	6.079	2
75)	16.82	15.51	0.8786	6.017	3.486	4.004	5.841	2
76)	16.77	15.62	0.8638	5.927	3.438	4.92	5.795	2
77)	17.32	15.91	0.8599	6.064	3.403	3.824	5.922	2
78)	20.71	17.23	0.8763	6.579	3.814	4.451	6.451	2
79)	18.94	16.49	0.875	6.445	3.639	5.064	6.362	2
80)	17.12	15.55	0.8892	5.85	3.566	2.858	5.746	2
81)	16.53	15.34	0.8823	5.875	3.467	5.532	5.88	2
82)	18.72	16.19	0.8977	6.006	3.857	5.324	5.879	2
83)	20.2	16.89	0.8894	6.285	3.864	5.173	6.187	2
84)	19.57	16.74	0.8779	6.384	3.772	1.472	6.273	2
85)	19.51	16.71	0.878	6.366	3.801	2.962	6.185	2
86)	18.27	16.09	0.887	6.173	3.651	2.443	6.197	2
87)	18.88	16.26	0.8969	6.084	3.764	1.649	6.109	2
88)	18.98	16.66	0.859	6.549	3.67	3.691	6.498	2
89)	21.18	17.21	0.8989	6.573	4.033	5.78	6.231	2
90)	20.88	17.05	0.9031	6.45	4.032	5.016	6.321	2
91)	20.1	16.99	0.8746	6.581	3.785	1.955	6.449	2
92)	18.76	16.2	0.8984	6.172	3.796	3.12	6.053	2
93)	18.81	16.29	0.8906	6.272	3.693	3.237	6.053	2
94)	18.59	16.05	0.9066	6.037	3.86	6.001	5.877	2
95)	18.36	16.52	0.8452	6.666	3.485	4.933	6.448	2
96)	16.87	15.65	0.8648	6.139	3.463	3.696	5.967	2
97)	19.31	16.59	0.8815	6.341	3.81	3.477	6.238	2
98)	18.98	16.57	0.8687	6.449	3.552	2.144	6.453	2
99)	18.17	16.26	0.8637	6.271	3.512	2.853	6.273	2
100)	18.72	16.34	0.881	6.219	3.684	2.188	6.097	2
101)	16.41	15.25	0.8866	5.718	3.525	4.217	5.618	2
102)	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	2
103)	19.46	16.5	0.8985	6.113	3.892	4.308	6.009	2
104)	19.18	16.63	0.8717	6.369	3.681	3.357	6.229	2
105)	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	2
106)	18.83	16.29	0.8917	6.037	3.786	2.553	5.879	2
107)	18.85	16.17	0.9056	6.152	3.806	2.843	6.2	2
108)	17.63	15.86	0.88	6.033	3.573	3.747	5.929	2

109)	19.94	16.92	0.8752	6.675	3.763	3.252	6.55	2
110)	18.55	16.22	0.8865	6.153	3.674	1.738	5.894	2
111)	18.45	16.12	0.8921	6.107	3.769	2.235	5.794	2
112)	19.38	16.72	0.8716	6.303	3.791	3.678	5.965	2
113)	19.13	16.31	0.9035	6.183	3.902	2.109	5.924	2
114)	19.14	16.61	0.8722	6.259	3.737	6.682	6.053	2
115)	20.97	17.25	0.8859	6.563	3.991	4.677	6.316	2
116)	19.06	16.45	0.8854	6.416	3.719	2.248	6.163	2
117)	18.96	16.2	0.9077	6.051	3.897	4.334	5.75	2
118)	19.15	16.45	0.889	6.245	3.815	3.084	6.185	2
119)	18.89	16.23	0.9008	6.227	3.769	3.639	5.966	2
120)	20.03	16.9	0.8811	6.493	3.857	3.063	6.32	2
121)	20.24	16.91	0.8897	6.315	3.962	5.901	6.188	2
122)	18.14	16.12	0.8772	6.059	3.563	3.619	6.011	2
123)	16.17	15.38	0.8588	5.762	3.387	4.286	5.703	2
124)	18.43	15.97	0.9077	5.98	3.771	2.984	5.905	2
125)	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
126)	18.75	16.18	0.8999	6.111	3.869	4.188	5.992	2
127)	18.65	16.41	0.8698	6.285	3.594	4.391	6.102	2
128)	17.98	15.85	0.8993	5.979	3.687	2.257	5.919	2
129)	20.16	17.03	0.8735	6.513	3.773	1.91	6.185	2
130)	17.55	15.66	0.8991	5.791	3.69	5.366	5.661	2
131)	18.3	15.89	0.9108	5.979	3.755	2.837	5.962	2
132)	18.94	16.32	0.8942	6.144	3.825	2.908	5.949	2
133)	15.38	14.9	0.8706	5.884	3.268	4.462	5.795	2
134)	16.16	15.33	0.8644	5.845	3.395	4.266	5.795	2
135)	15.56	14.89	0.8823	5.776	3.408	4.972	5.847	2
136)	15.38	14.66	0.899	5.477	3.465	3.6	5.439	2
137)	17.36	15.76	0.8785	6.145	3.574	3.526	5.971	2
138)	15.57	15.15	0.8527	5.92	3.231	2.64	5.879	2
139)	15.6	15.11	0.858	5.832	3.286	2.725	5.752	2
140)	16.23	15.18	0.885	5.872	3.472	3.769	5.922	2
141)	13.07	13.92	0.848	5.472	2.994	5.304	5.395	3
142)	13.32	13.94	0.8613	5.541	3.073	7.035	5.44	3
143)	13.34	13.95	0.862	5.389	3.074	5.995	5.307	3
144)	12.22	13.32	0.8652	5.224	2.967	5.469	5.221	3
145)	11.82	13.4	0.8274	5.314	2.777	4.471	5.178	3

146)	11.21	13.13	0.8167	5.279	2.687	6.169	5.275	3
147)	11.43	13.13	0.8335	5.176	2.719	2.221	5.132	3
148)	12.49	13.46	0.8658	5.267	2.967	4.421	5.002	3
149)	12.7	13.71	0.8491	5.386	2.911	3.26	5.316	3
150)	10.79	12.93	0.8107	5.317	2.648	5.462	5.194	3
151)	11.83	13.23	0.8496	5.263	2.84	5.195	5.307	3
152)	12.01	13.52	0.8249	5.405	2.776	6.992	5.27	3
153)	12.26	13.6	0.8333	5.408	2.833	4.756	5.36	3
154)	11.18	13.04	0.8266	5.22	2.693	3.332	5.001	3
155)	11.36	13.05	0.8382	5.175	2.755	4.048	5.263	3
156)	11.19	13.05	0.8253	5.25	2.675	5.813	5.219	3
157)	11.34	12.87	0.8596	5.053	2.849	3.347	5.003	3
158)	12.13	13.73	0.8081	5.394	2.745	4.825	5.22	3
159)	11.75	13.52	0.8082	5.444	2.678	4.378	5.31	3
160)	11.49	13.22	0.8263	5.304	2.695	5.388	5.31	3
161)	12.54	13.67	0.8425	5.451	2.879	3.082	5.491	3
162)	12.02	13.33	0.8503	5.35	2.81	4.271	5.308	3
163)	12.05	13.41	0.8416	5.267	2.847	4.988	5.046	3
164)	12.55	13.57	0.8558	5.333	2.968	4.419	5.176	3
165)	11.14	12.79	0.8558	5.011	2.794	6.388	5.049	3
166)	12.1	13.15	0.8793	5.105	2.941	2.201	5.056	3
167)	12.44	13.59	0.8462	5.319	2.897	4.924	5.27	3
168)	12.15	13.45	0.8443	5.417	2.837	3.638	5.338	3
169)	11.35	13.12	0.8291	5.176	2.668	4.337	5.132	3
170)	11.24	13	0.8359	5.09	2.715	3.521	5.088	3
171)	11.02	13	0.8189	5.325	2.701	6.735	5.163	3
172)	11.55	13.1	0.8455	5.167	2.845	6.715	4.956	3
173)	11.27	12.97	0.8419	5.088	2.763	4.309	5	3
174)	11.4	13.08	0.8375	5.136	2.763	5.588	5.089	3
175)	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	3
176)	10.8	12.57	0.859	4.981	2.821	4.773	5.063	3
177)	11.26	13.01	0.8355	5.186	2.71	5.335	5.092	3
178)	10.74	12.73	0.8329	5.145	2.642	4.702	4.963	3
179)	11.48	13.05	0.8473	5.18	2.758	5.876	5.002	3
180)	12.21	13.47	0.8453	5.357	2.893	1.661	5.178	3
181)	11.41	12.95	0.856	5.09	2.775	4.957	4.825	3
182)	12.46	13.41	0.8706	5.236	3.017	4.987	5.147	3

183)	12.19	13.36	0.8579	5.24	2.909	4.857	5.158	3
184)	11.65	13.07	0.8575	5.108	2.85	5.209	5.135	3
185)	12.89	13.77	0.8541	5.495	3.026	6.185	5.316	3
186)	11.56	13.31	0.8198	5.363	2.683	4.062	5.182	3
187)	11.81	13.45	0.8198	5.413	2.716	4.898	5.352	3
188)	10.91	12.8	0.8372	5.088	2.675	4.179	4.956	3
189)	11.23	12.82	0.8594	5.089	2.821	7.524	4.957	3
190)	10.59	12.41	0.8648	4.899	2.787	4.975	4.794	3
191)	10.93	12.8	0.839	5.046	2.717	5.398	5.045	3
192)	11.27	12.86	0.8563	5.091	2.804	3.985	5.001	3
193)	11.87	13.02	0.8795	5.132	2.953	3.597	5.132	3
194)	10.82	12.83	0.8256	5.18	2.63	4.853	5.089	3
195)	12.11	13.27	0.8639	5.236	2.975	4.132	5.012	3
196)	12.8	13.47	0.886	5.16	3.126	4.873	4.914	3
197)	12.79	13.53	0.8786	5.224	3.054	5.483	4.958	3
198)	13.37	13.78	0.8849	5.32	3.128	4.67	5.091	3
199)	12.62	13.67	0.8481	5.41	2.911	3.306	5.231	3
200)	12.76	13.38	0.8964	5.073	3.155	2.828	4.83	3
201)	12.38	13.44	0.8609	5.219	2.989	5.472	5.045	3
202)	12.67	13.32	0.8977	4.984	3.135	2.3	4.745	3
203)	11.18	12.72	0.868	5.009	2.81	4.051	4.828	3
204)	12.7	13.41	0.8874	5.183	3.091	8.456	5	3
205)	12.37	13.47	0.8567	5.204	2.96	3.919	5.001	3
206)	12.19	13.2	0.8783	5.137	2.981	3.631	4.87	3
207)	11.23	12.88	0.8511	5.14	2.795	4.325	5.003	3
208)	13.2	13.66	0.8883	5.236	3.232	8.315	5.056	3
209)	11.84	13.21	0.8521	5.175	2.836	3.598	5.044	3
210)	12.3	13.34	0.8684	5.243	2.974	5.637	5.063	3

Compactness به این صورت تعریف شده است:

$$C = \frac{4\pi A}{p^2}$$

$$\pi = 3.14$$

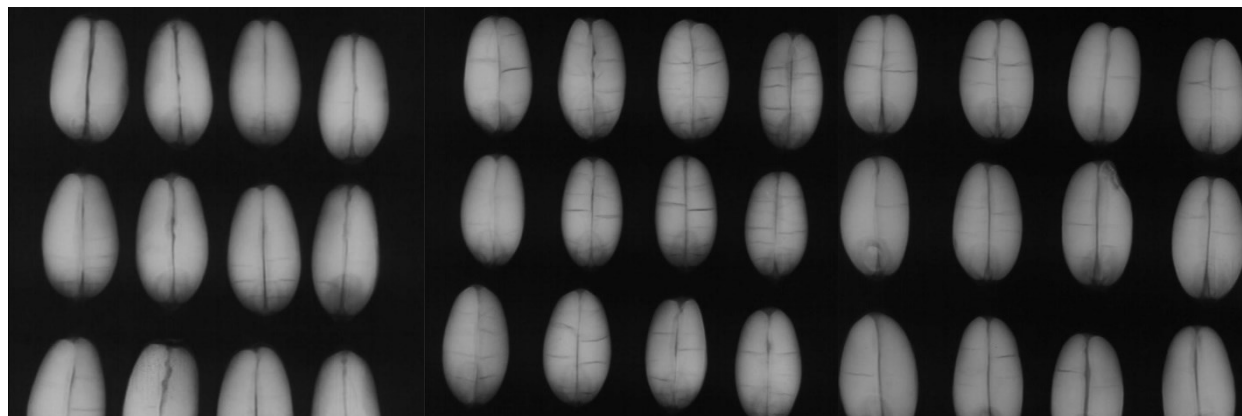
p = perimeter محیط

(میلی متر)

a = Area مساحت

(میلی متر مربع)

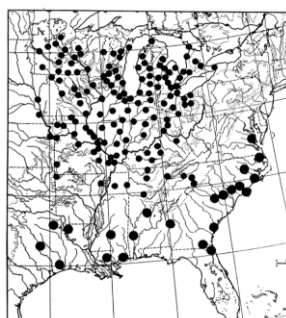
فتوگرام‌های اکس-ری نمونه از سه دسته دانه‌ی گندم بررسی شده



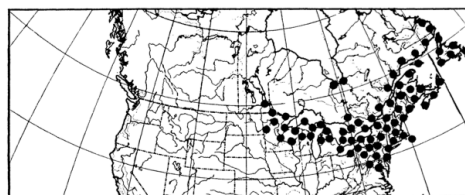
Canadian

Kama

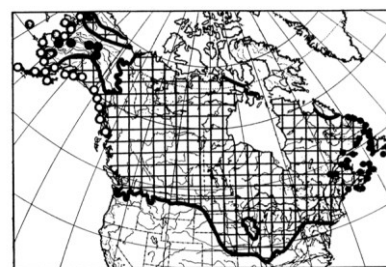
Rosa



Map 3. Range of *Iris virginica* (large circles), and of *I. virginica* var. *Abrexi* (small circles).



Map 2. Range of *Iris versicolor*.



Map 1. Range of *Iris setosa* (open circles), *I. setosa* var. *canadensis* (small solid circles), and *I. setosa* var. *interior* (large solid circles). Cross hatching shows extent of maximum Pleistocene glaciation.

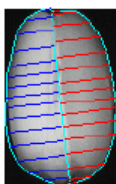
در بالا نقشه پراکندگی هریک از این سه گونه در ایالات متحده مشاهده می‌شود.

File name with path: D:\RTG_IM\1\RTG42.BMP

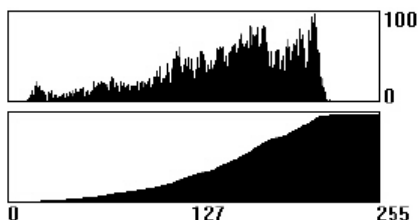
Projection: A

Label: rtg42, g1

Comments: var. Rosa



Area: 20.32
Perimeter: 16.8
Compactness: 0.9047
Length: 6.344
Width: 4.017
Asymmetry: 2.782
Groove : 6.052



Median: 158
Mode: 211
Average brightness: 147
Std. dev.: 48.78
Min: 12 Max: 222
Area [hist]: 20.33

نمونه سنجش دانه‌ای از

دسته‌ی Rosa:

برای اندازه‌گیری هر دانه در قبال
تکنولوژی لیزر یا میکروسکوپی از
تکنیک اشعه اکس ملایم استفاده
شده است.

توصیف آماری پایه‌ای داده - اندازه‌گیری گرایش به مرکز

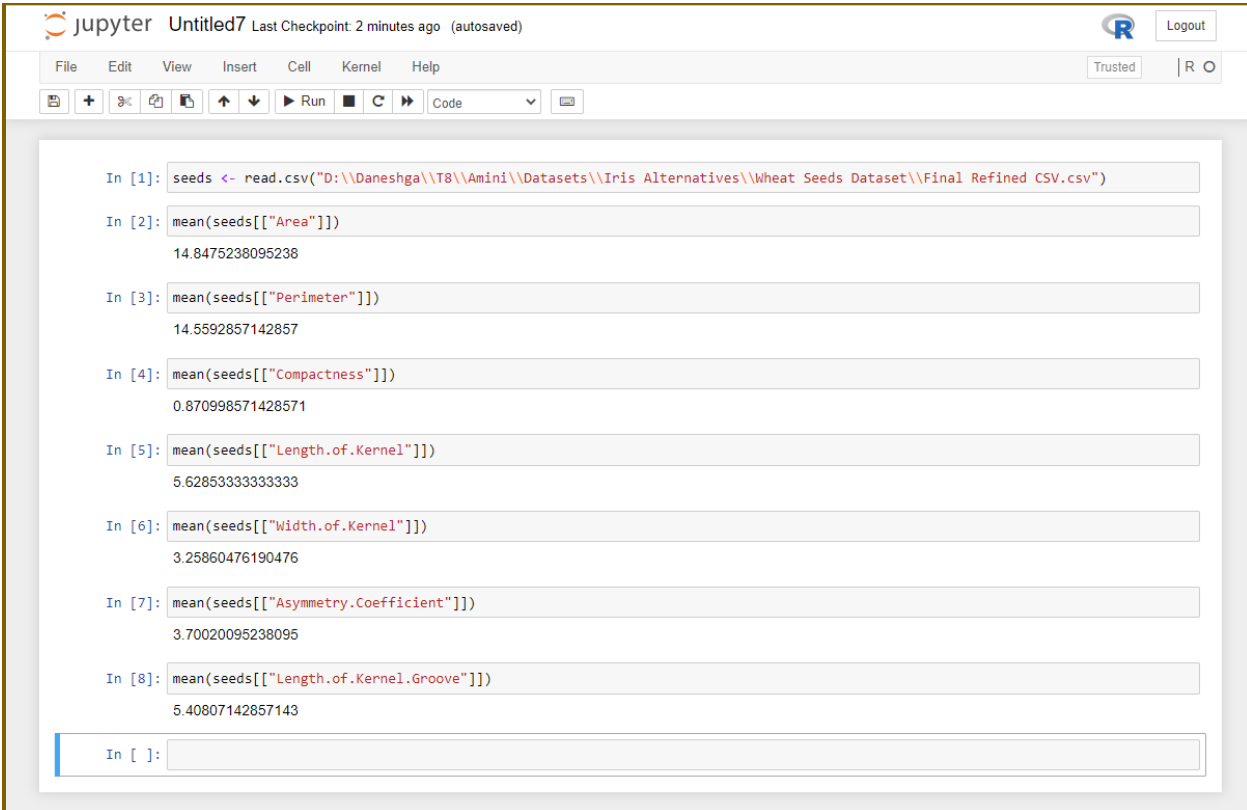
برای محاسبات از محیط Jupyter Notebook و زبان R استفاده شده است.

Mean

میانگین

پس از تبدیل فرمت فایل اولیه به **csv**. ([صفحه ۴](#))، اضافه کردن عنوان هر ویژگی به ستون مربوطه و اضافه کردن ستونی جدید تحت عنوان **Instances**، فایل نهایی را با دستور **read.csv** در محیط ایمپورت کرده و آن را به عنوان **seeds** نسبت دادم.

سپس با دستور **mean(seeds[["Column Title"]])** میانگین را برای هر ستون (جز دو ستون کلاس و Instance) محاسبه کردم.



```
Jupyter Untitled7 Last Checkpoint: 2 minutes ago (autosaved)
File Edit View Insert Cell Kernel Help Trusted | R
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds Dataset\\Final Refined CSV.csv")
In [2]: mean(seeds[["Area"]])
14.8475238095238
In [3]: mean(seeds[["Perimeter"]])
14.5592857142857
In [4]: mean(seeds[["Compactness"]])
0.870998571428571
In [5]: mean(seeds[["Length.of.Kernel"]])
5.62853333333333
In [6]: mean(seeds[["Width.of.Kernel"]])
3.25860476190476
In [7]: mean(seeds[["Asymmetry.Coefficient"]])
3.70020095238095
In [8]: mean(seeds[["Length.of.Kernel.Groove"]])
5.40807142857143
In [ ]:
```

با دستور **summary** می‌توان مینیمم، ماکسیمم، میانگین، میانه (چارک دوم)، چارک اول و چارک سوم را در R برای هر فیلد محاسبه کرد.

با قطعه کد زیر دستور **summary** را پس از تفکیک ستون‌های عددی (برای پرهیز از مقادیر بی‌معنی) اجرا کردم.

```
cols <- c("Area", "Perimeter", "Compactness", "Length.of.Kernel",  
"Width.of.Kernel", "Asymmetry.Coefficient",  
"Length.of.Kernel.Groove");summary(seeds[cols])
```

```
In [10]: cols <- c("Area", "Perimeter", "Compactness", "Length.of.Kernel", "W
```

Area	Perimeter	Compactness	Length.of.Kernel
Min. :10.59	Min. :12.41	Min. :0.8081	Min. :4.899
1st Qu.:12.27	1st Qu.:13.45	1st Qu.:0.8569	1st Qu.:5.262
Median :14.36	Median :14.32	Median :0.8734	Median :5.524
Mean :14.85	Mean :14.56	Mean :0.8710	Mean :5.629
3rd Qu.:17.30	3rd Qu.:15.71	3rd Qu.:0.8878	3rd Qu.:5.980
Max. :21.18	Max. :17.25	Max. :0.9183	Max. :6.675

Width.of.Kernel	Asymmetry.Coefficient	Length.of.Kernel.Groove
Min. :2.630	Min. :0.7651	Min. :4.519
1st Qu.:2.944	1st Qu.:2.5615	1st Qu.:5.045
Median :3.237	Median :3.5990	Median :5.223
Mean :3.259	Mean :3.7002	Mean :5.408
3rd Qu.:3.562	3rd Qu.:4.7687	3rd Qu.:5.877
Max. :4.033	Max. :8.4560	Max. :6.550

Median

علاوہ بر محاسبہ‌ی میانہ به صورت کلی در صفحہ‌ی قبل (با دستور summary)، آن را به صورت تکی برای هر فیلد، با دستور `median(seeds[,Column Num])` که در آن **Column Num** شماره هر ستون در صورت شماره گذاری از چپ به راست می‌باشد را حساب کردم.

شماره ستون‌ها برای در نظر نگرفتن ستون‌های **instance** (ستون اول) و **seedtype** (ستون نهم) از ۲ شروع و به ۸ ختم می‌شوند.

```
In [12]: median(seeds[,2])
```

14.355

```
In [13]: median(seeds[,3])
```

14.32

```
In [14]: median(seeds[,4])
```

0.87345

```
In [15]: median(seeds[,5])
```

5.5235

```
In [16]: median(seeds[,6])
```

3.237

```
In [17]: median(seeds[,7])
```

3.599

```
In [18]: median(seeds[,8])
```

5.223

Mode

R برای محاسبه مد عملگری استاندارد و **in-built** (از پیش تعریف شده) ندارد. بنابراین یک تابع تعریف می‌کنیم تا مد یک مجموعه داده را در R محاسبه کند. این تابع برداری را به عنوان ورودی می‌گیرد و مقدار مد را به عنوان خروجی تحویل می‌دهد.

```
Mode <- function(x) {
  unique_x <- unique(x)
  unique_x [which.max(tabulate(match(x, unique_x)))]
}
```

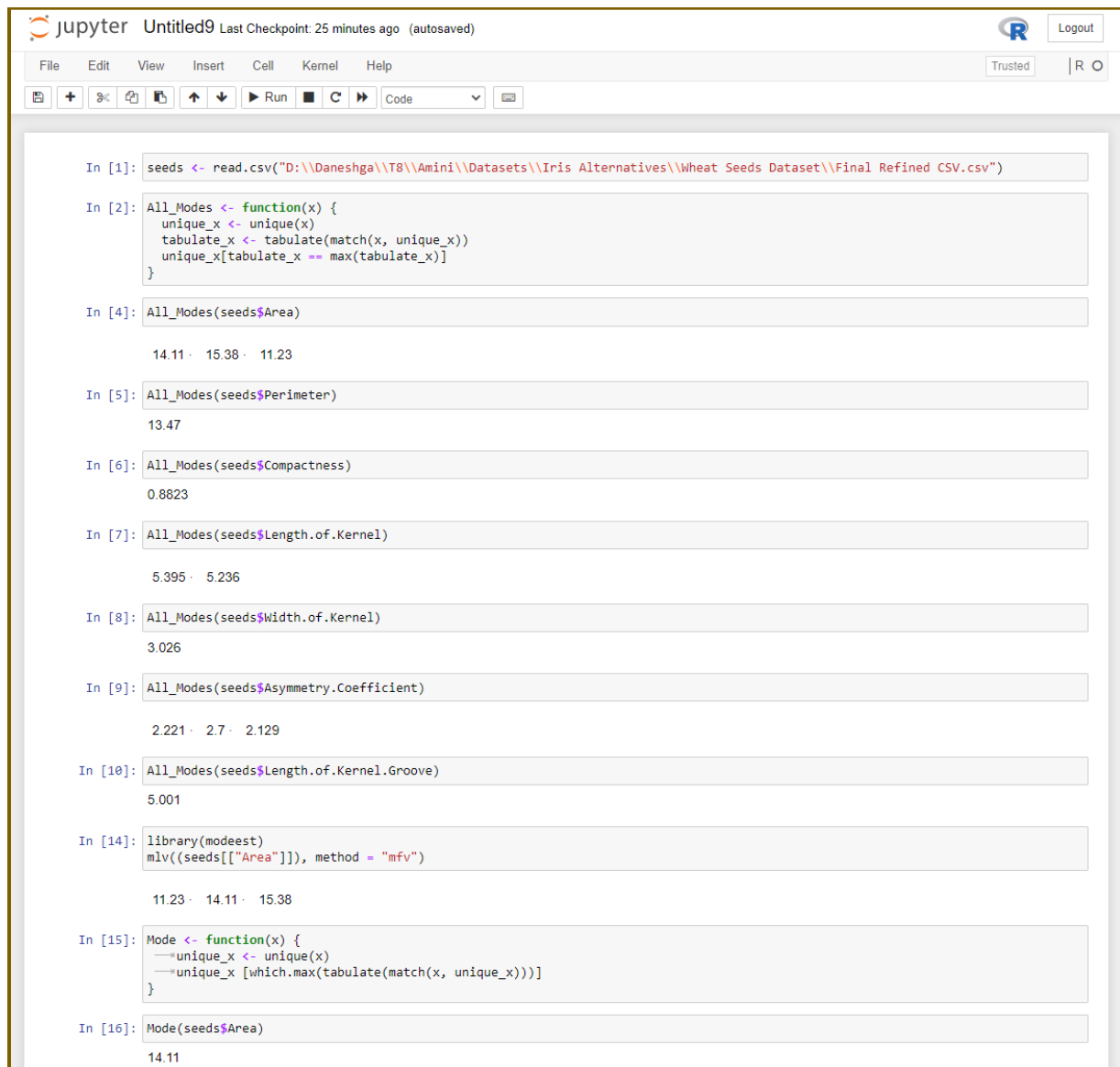
کد بالا در صورت وجود بیش از یک مد در مجموعه داده‌ی ما (**multimodal** بودن)، مانند دستور **which.max** عمل می‌کند و فقط اولین مقداری که برای مد پیدا شد را بر می‌گرداند. برای رفع این مشکل با کد زیر تابع را تعریف می‌کنیم.

```
All_Modes <- function(x) {
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}
```

پس از تعریف تابع، با فرمان **All_Modes(seeds\$Column Title)** می‌توان مد (ها) را برای هر ستون مشخص کرد.

- علاوه بر استفاده از تابع بالا برای محاسبه مد در R می‌توان از کتابخانه‌ی **modeest (mode estimation)** نیز بهره برد، پس از آخرین محاسبه، به عنوان نمونه با استفاده از دو فرمان این کتابخانه را به کار گرفتیم.

- همچنین تابع **unimodal** صفحه‌ی گذشته در آخرین نمونه تعریف و برای ستون **trimodal** مساحت استفاده شده، همانطور که مشاهده می‌شود، این تابع به اولین مقدار برای مد بسنده می‌کند.



```
jupyter Untitled9 Last Checkpoint: 25 minutes ago (autosaved)
File Edit View Insert Cell Kernel Help Trusted | R O

In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds Dataset\\Final Refined CSV.csv")

In [2]: All_Modes <- function(x) {
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}

In [4]: All_Modes(seeds$Area)

14.11 · 15.38 · 11.23

In [5]: All_Modes(seeds$Perimeter)

13.47

In [6]: All_Modes(seeds$Compactness)

0.8823

In [7]: All_Modes(seeds$Length.of.Kernel)

5.395 · 5.236

In [8]: All_Modes(seeds$Width.of.Kernel)

3.026

In [9]: All_Modes(seeds$Asymmetry.Coefficient)

2.221 · 2.7 · 2.129

In [10]: All_Modes(seeds$Length.of.Kernel.Groove)

5.001

In [14]: library(modeest)
mlv((seeds[["Area"]]), method = "mfv")

11.23 · 14.11 · 15.38

In [15]: Mode <- function(x) {
  →unique_x <- unique(x)
  →unique_x [which.max(tabulate(match(x, unique_x)))]
}

In [16]: Mode(seeds$Area)

14.11
```


همانطور که مشاهده می‌شود، فیلد **Area** با سه تکرار برای هر مقدار و فیلد **Asymmetry Coefficient** با دو تکرار برای هر مقدار، **Trimodal** تشخیص داده شدند.

فیلد **Length of Kernel** با سه تکرار، **bimodal** و فیلدهای دیگر هم با مدهایی با تعداد تکرار متفاوت در دیتاست **unimodal** هستند.

(برای نصب کتابخانه **modeest** از فرمان زیر در ترمینال مامبا استفاده شد.)
`mamba install -c conda-forge r-modeest`

Midrange

محدوده میانی

متاسفانه R برای محاسبه‌ی محدوده میانی نیز دستوری از پیش تعریف شده ندارد گرچه می‌توان به طور دستی از مقادیر Min و Max که توسط فرمان summary برای هر ستون بدست آوردیم استفاده کنیم.

با نصب کتابخانه directlabels یا statip می‌توان از دستور midrange برای محاسبه آن برای هر ویژگی استفاده کرد.

(برای نصب کتابخانه directlabels از فرمان زیر در ترمینال مامبا استفاده شد.)

```
mamba install -c conda-forge r-directlabels
```

(برای نصب کتابخانه statip نیز می‌توان از فرمان زیر استفاده کرد.)

```
mamba install -c conda-forge static
```

پس از اینکلود کردن کتابخانه به طریق زیر میتوان محدوده میانی را برای هر ستون محاسبه کرد.

```
library(directlabels)
```

```
midrange(seeds$Column Title)
```

```
In [2]: library(statip)
```

```
In [3]: midrange(seeds$Area)
midrange(seeds$Perimeter)
midrange(seeds$Compactness)
midrange(seeds$Length.of.Kernel)
midrange(seeds$Width.of.Kernel)
midrange(seeds$Asymmetry.Coefficient)
midrange(seeds$Length.of.Kernel.Groove)
```

15.885

14.83


0.8632

5.787


3.3315

4.61055

5.5345











 jupyter

Untitled12


 Logout

FileEditViewInsertCellKernelHelp

Trusted | R



Code



```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\W")
In [2]: library(directlabels)
In [3]: midrange(seeds$Area)
15.885
In [4]: midrange(seeds$Perimeter)
14.83
In [5]: midrange(seeds$Compactness)
0.8632
In [6]: midrange(seeds$Length.of.Kernel)
5.787
In [7]: midrange(seeds$Width.of.Kernel)
3.3315
In [8]: midrange(seeds$Asymmetry.Coefficient)
4.61055
In [9]: midrange(seeds$Length.of.Kernel.Groove)
5.5345
```

با مقایسه مد و میانه در تصویر پایین متوجه می‌شویم که ویژگی‌های محیط، طول هسته، پهنای هسته، ضریب نامتقارنی و طول groove دانه **اریب مثبت (Positively Skewed)** بوده و **Compactness**، **اریب منفی (Negatively Skewed)** می‌باشد. مدهای فیلد مساحت هم از میانه بزرگ‌تر و هم کوچک‌تر هستند (درمورد این حالت بحث نشد).

```
In [3]: median(seeds[,2])  
All_Modes(seeds$Area)  
  
14.355
```

14.11 · 15.38 · 11.23

```
In [4]: median(seeds[,3])  
All_Modes(seeds$Perimeter)  
  
14.32  
13.47
```

```
In [5]: median(seeds[,4])  
All_Modes(seeds$Compactness)  
  
0.87345  
0.8823
```

```
In [6]: median(seeds[,5])  
All_Modes(seeds$Length.of.Kernel)  
  
5.5235  
  
5.395 · 5.236
```

```
In [7]: median(seeds[,6])  
All_Modes(seeds$Width.of.Kernel)  
  
3.237  
3.026
```

```
In [8]: median(seeds[,7])  
All_Modes(seeds$Asymmetry.Coefficient)  
  
3.599  
  
2.221 · 2.7 · 2.129
```

```
In [9]: median(seeds[,8])  
All_Modes(seeds$Length.of.Kernel.Groove)  
  
5.223  
5.001
```

توصیف آماری پایه‌ای داده - اندازه‌گیری پراکندگی داده‌ها

Range

برد

برد تفاوت بین بزرگترین مقدار و کوچکترین مقدار در مجموعه است. برای محاسبه آن می‌توان به طور دستی از مقادیر Min و Max که توسط فرمان summary برای هر ستون بدست آوردیم استفاده کرد.

همچنین دستور Range در R (که در این مثال برای داده‌ای تحت عنوان "seeds" با سینتکس (range(seeds\$Column Title, na.rm=TRUE) فقط مقادیر Max و Min را برای ستون (صفت) مشخص شده به ما تحویل می‌دهد.

برای اینکه داده‌ای را به صورت دستی وارد نکنیم از کد زیر (با محاسبه Min و Max برای ویژگی و تفریق آنان) برد را برای هر ویژگی بدست آوردم.

max(seeds\$Column Title, na.rm=TRUE) -

min(seeds\$Column Title, na.rm=TRUE)

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives
```

```
In [2]: range(seeds$Area, na.rm=TRUE)
```


```
10.59 - 21.18
```

```
In [4]: 21.18 - 10.59
```










```
10.59
```

```
In [5]: max(seeds$Area, na.rm=TRUE) - min(seeds$Area, na.rm=TRUE)
```

```
10.59
```

Jupyter Untitled19 Last Checkpoint: 3 minutes ago (unsaved changes)  Logout

File Edit View Insert Cell Kernel Help Trusted | R

         Code

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds Dataset\\Final Refined CSV.csv")

In [2]: max(seeds$Area, na.rm=TRUE) - min(seeds$Area, na.rm=TRUE)
10.59

In [3]: max(seeds$Perimeter, na.rm=TRUE) - min(seeds$Perimeter, na.rm=TRUE)
4.84

In [4]: max(seeds$Compactness, na.rm=TRUE) - min(seeds$Compactness, na.rm=TRUE)
0.1102

In [5]: max(seeds$Length.of.Kernel, na.rm=TRUE) - min(seeds$Length.of.Kernel, na.rm=TRUE)
1.776

In [6]: max(seeds$Width.of.Kernel, na.rm=TRUE) - min(seeds$Width.of.Kernel, na.rm=TRUE)
1.403

In [7]: max(seeds$Asymmetry.Coefficient, na.rm=TRUE) - min(seeds$Asymmetry.Coefficient, na.rm=TRUE)
7.6909

In [8]: max(seeds$Length.of.Kernel.Groove, na.rm=TRUE) - min(seeds$Length.of.Kernel.Groove, na.rm=TRUE)
2.031

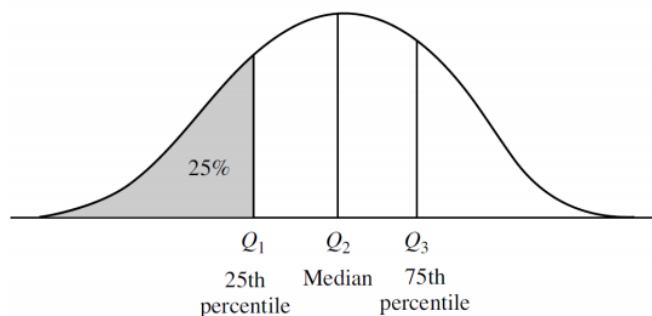
In [ ]:
```

Quantiles

میان، چارک و صدک

علاوه بر روش زیرین (فرمان quantile)، میان در صفحه‌ی ۱۴ و چارک‌ها در صفحه‌ی ۱۳ (با دستور summary) برای تمام ویژگی‌ها محاسبه شده‌است.

چارک و میان: در R تابع quantile پنج مقدار را به عنوان خروجی به ما تحویل می‌دهد که در آن مقادیر 0% و 100% حداقل و حداکثر، و سه مقدار دیگر 25%، 50% و 75% به ترتیب چارک اول (صدک بیست و پنجم)، دوم (میان/ صدک پنجاهم) و سوم (صدک هفتاد و پنجم) هستند.



```
In [8]: quantile (seeds$Area)
quantile (seeds$Perimeter)
quantile (seeds$Compactness)
quantile (seeds$Length.of.Kernel)
quantile (seeds$Width.of.Kernel)
quantile (seeds$Asymmetry.Coefficient)
quantile (seeds$Length.of.Kernel.Groove)|
```

0%: 10.59 25%: 12.27 50%: 14.355 75%: 17.305 100%: 21.18

0%: 12.41 25%: 13.45 50%: 14.32 75%: 15.715 100%: 17.25

0%: 0.8081 25%: 0.8569 50%: 0.87345 75%: 0.887775 100%: 0.9183

0%: 4.899 25%: 5.26225 50%: 5.5235 75%: 5.97975 100%: 6.675

0%: 2.63 25%: 2.944 50%: 3.237 75%: 3.56175 100%: 4.033

0%: 0.7651 25%: 2.5615 50%: 3.599 75%: 4.76875 100%: 8.456

0%: 4.519 25%: 5.045 50%: 5.223 75%: 5.877 100%: 6.55

صدک: با استفاده از تابع `quantile` نیز می‌توان صدک‌های مورد نظر را پیدا کرد، به عنوان مثال، در تصویر زیر صدک سی‌ودوم، پنجاه‌وهفتم و نودوهشتم در قالب سه مقدار به عنوان خروجی به ما تحویل داده شد.

```
In [11]: quantile(seeds$Area, c(.32, .57, .98))
```

```
32%: 12.7288 57%: 14.9926 98%: 20.2328
```

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds Data.csv")
```

```
In [13]: quantile(seeds$Area, c(.40, .57, .80))
quantile(seeds$Perimeter, c(.11, .68, .98))
quantile(seeds$Compactness, c(.32, .43, .23))
quantile(seeds$Length.of.Kernel, c(.96, .53, .13))
quantile(seeds$Width.of.Kernel, c(.43, .87, .76))
quantile(seeds$Asymmetry.Coefficient, c(.78, .55, .28))
quantile(seeds$Length.of.Kernel.Groove, c(.64, .62, .33))
```

```
40%: 13.418 57%: 14.9926 80%: 18.312
```

```
11%: 13.0499 68%: 15.1884 98%: 17.0228
```

```
32%: 0.862264 43%: 0.869483 23%: 0.855814
```

```
96%: 6.47752 53%: 5.56762 13%: 5.14738
```

```
43%: 3.15587 87%: 3.76815 76%: 3.56552
```

```
78%: 4.92008 55%: 3.82125 28%: 2.70208
```

```
64%: 5.44228 62%: 5.39558 33%: 5.091
```

```
In [12]: quantile(seeds$Perimeter, c(0.125, 0.375, 0.625, 0.875))
```

```
12.5%: 13.07125 37.5%: 13.795 62.5%: 14.87875 87.5%: 16.31875
```

در تصویر بالا، برای صفت‌های متفاوت صدک‌های مختلف محاسبه شد و در آخرین `cell` مشاهده می‌شود که در صورت نیاز می‌توانیم از صدک هم فراتر رفته و «هزارک» برای ویژگی‌های خود محاسبه کنیم.

IQR (Interquartile Range)

محدوده میان ربعی

می‌توانیم با تابع `quantile` در R، چارک اول را از چارک سوم کم کنیم (همان‌طور که برای فیلد `Area` نمایش داده شده است).

اما R برای محاسبه محدوده میان ربعی تابع `IQR` را به‌طور پیش‌فرض در اختیار ما قرار داده است که در ادامه با استفاده از آن این مقادیر را برای ویژگی‌ها محاسبه کردم.

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat :
```

```
In [16]: quantile(seeds$Area, c(.75)) - quantile(seeds$Area, c(.25))  
IQR(seeds$Area)
```

```
75%: 5.035
```

```
5.035
```

```
In [17]: IQR(seeds$Perimeter)  
IQR(seeds$Compactness)  
IQR(seeds$Length.of.Kernel)  
IQR(seeds$Width.of.Kernel)  
IQR(seeds$Asymmetry.Coefficient)  
IQR(seeds$Length.of.Kernel.Groove)
```

```
2.265
```

```
0.030875
```

```
0.7175
```

```
0.61775
```

```
2.20725
```

```
0.832
```

Five-Number Summary

خلاصه پنج عددی

خلاصه پنج عددی به مجموعه پنج **quantile** گفته می‌شود که خلاصه‌ای مناسب را از داده به ما ارائه می‌دهد. مینیمم، چارک اول، دوم (میانه)، سوم و ماکسیمم.

تمامی این مقادیر در گذشته و همچنین با دستور **summary** در صفحه‌ی ۱۳ برای تمام ویژگی‌ها محاسبه شده‌اند.

R قابلیت محاسبه خلاصه پنج عددی را با فانکشن **fivenum** به صورت زیر فراهم می‌سازد.

fivenum(seeds\$Column Title)

```
jupyter Untitled20 Last Checkpoint: an hour ago (unsaved changes)
File Edit View Insert Cell Kernel Help Trusted | R O

In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds Dataset\\Final Refined CSV.csv")

In [24]: #<code>fivenum(seeds$Area)</code>
fivenum(seeds$Area)
fivenum(seeds$Perimeter)
fivenum(seeds$Compactness)
fivenum(seeds$Length.of.Kernel)
fivenum(seeds$Width.of.Kernel)
fivenum(seeds$Asymmetry.Coefficient)
fivenum(seeds$Length.of.Kernel.Groove)
cols <- c("Area", "Perimeter", "Compactness", "Length.of.Kernel", "Width.of.Kernel", "Asymmetry.Coefficient", "Length.of.Kernel.Groove")

10.59 · 12.26 · 14.355 · 17.32 · 21.18

12.41 · 13.45 · 14.32 · 15.73 · 17.25

0.8081 · 0.8567 · 0.87345 · 0.8879 · 0.9183

4.899 · 5.262 · 5.5235 · 5.98 · 6.675

2.63 · 2.941 · 3.237 · 3.562 · 4.033

0.7651 · 2.553 · 3.599 · 4.773 · 8.456

4.519 · 5.045 · 5.223 · 5.877 · 6.55

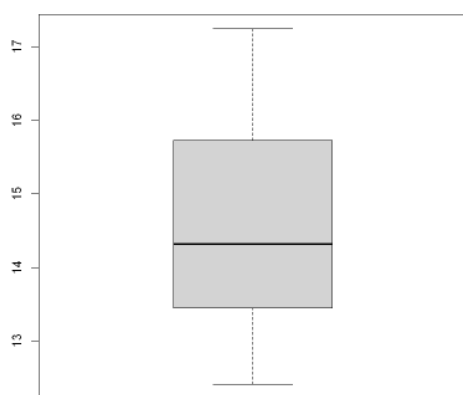
Area      Perimeter  Compactness  Length.of.Kernel
Min. :10.59  Min. :12.41  Min. :0.8081  Min. :4.899
1st Qu.:12.27 1st Qu.:13.45 1st Qu.:0.8569 1st Qu.:5.262
Median :14.36 Median :14.32 Median :0.8734 Median :5.524
Mean :14.85  Mean :14.56  Mean :0.8710  Mean :5.629
3rd Qu.:17.30 3rd Qu.:15.71 3rd Qu.:0.8878 3rd Qu.:5.980
Max. :21.18  Max. :17.25  Max. :0.9183  Max. :6.675
Width.of.Kernel Asymmetry.Coefficient Length.of.Kernel.Groove
Min. :2.630  Min. :0.7651  Min. :4.519
1st Qu.:2.944 1st Qu.:2.5615 1st Qu.:5.045
Median :3.237 Median :3.5990 Median :5.223
Mean :3.259  Mean :3.7002  Mean :5.408
3rd Qu.:3.562 3rd Qu.:4.7687 3rd Qu.:5.877
Max. :4.033  Max. :8.4560  Max. :6.550
```

Boxplot

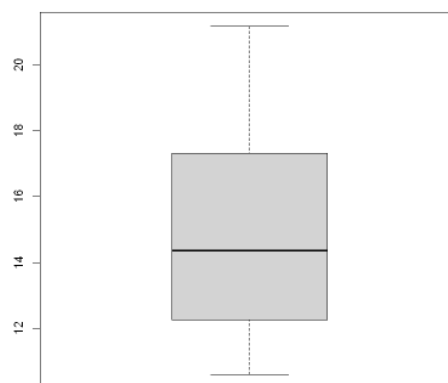
نمودار جعبه‌ای

خوشبختانه R برای ترسیم **boxplot** دستوری ساده و با همین نام دارد که در ادامه استفاده از آن مشاهده می‌شود.

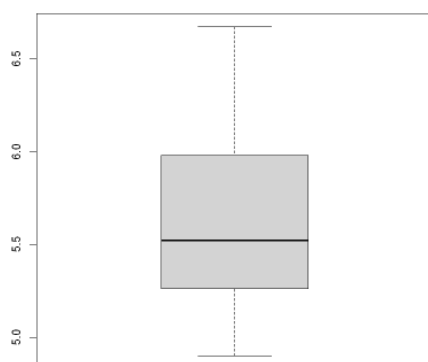
```
In [27]: boxplot(seeds$Perimeter)
```



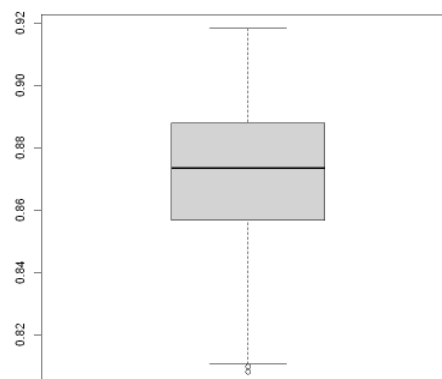
```
In [26]: boxplot(seeds$Area)
```



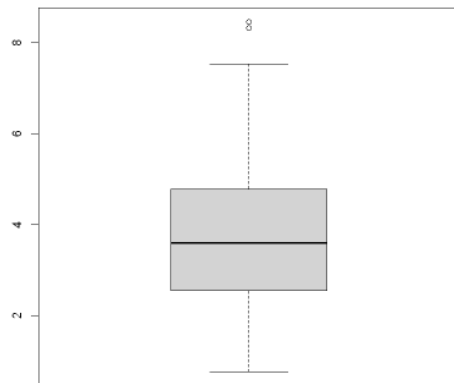
```
In [29]: boxplot(seeds$Length.of.Kernel)
```



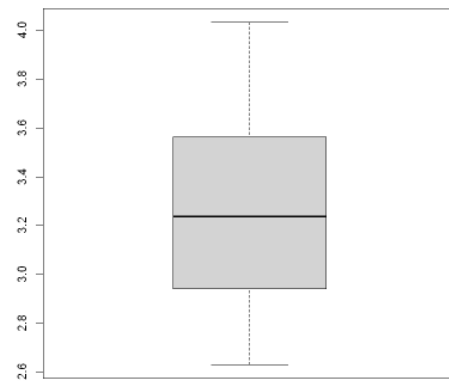
```
In [28]: boxplot(seeds$Compactness)
```



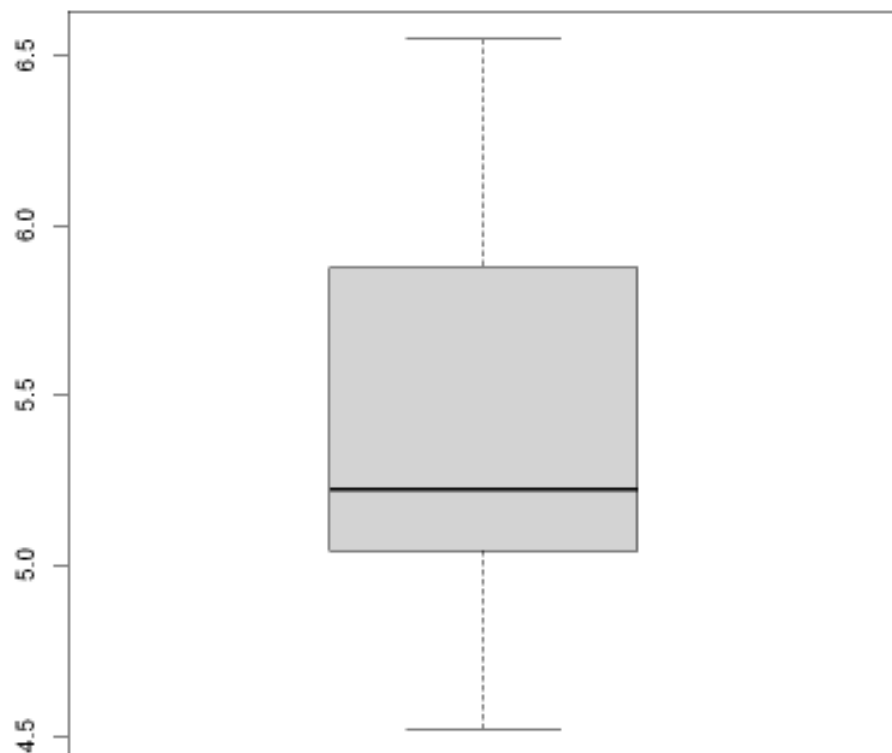
```
In [31]: boxplot(seeds$Asymmetry.Coefficient)
```



```
In [30]: boxplot(seeds$width.of.Kernel)
```

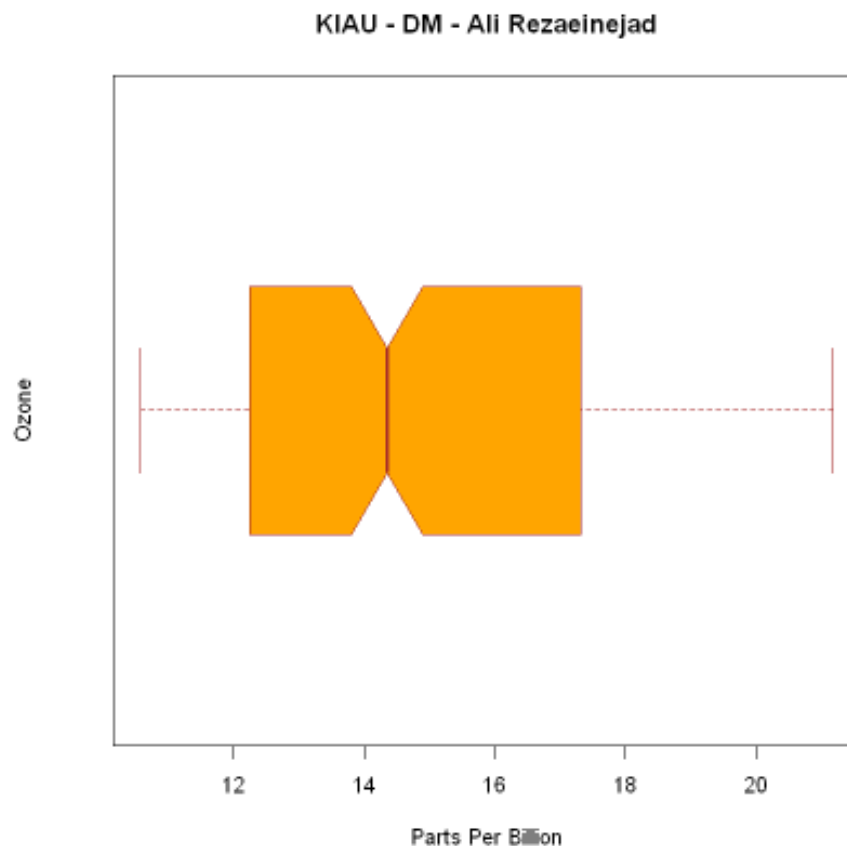


```
In [32]: boxplot(seeds$Length.of.Kernel.Groove)
```



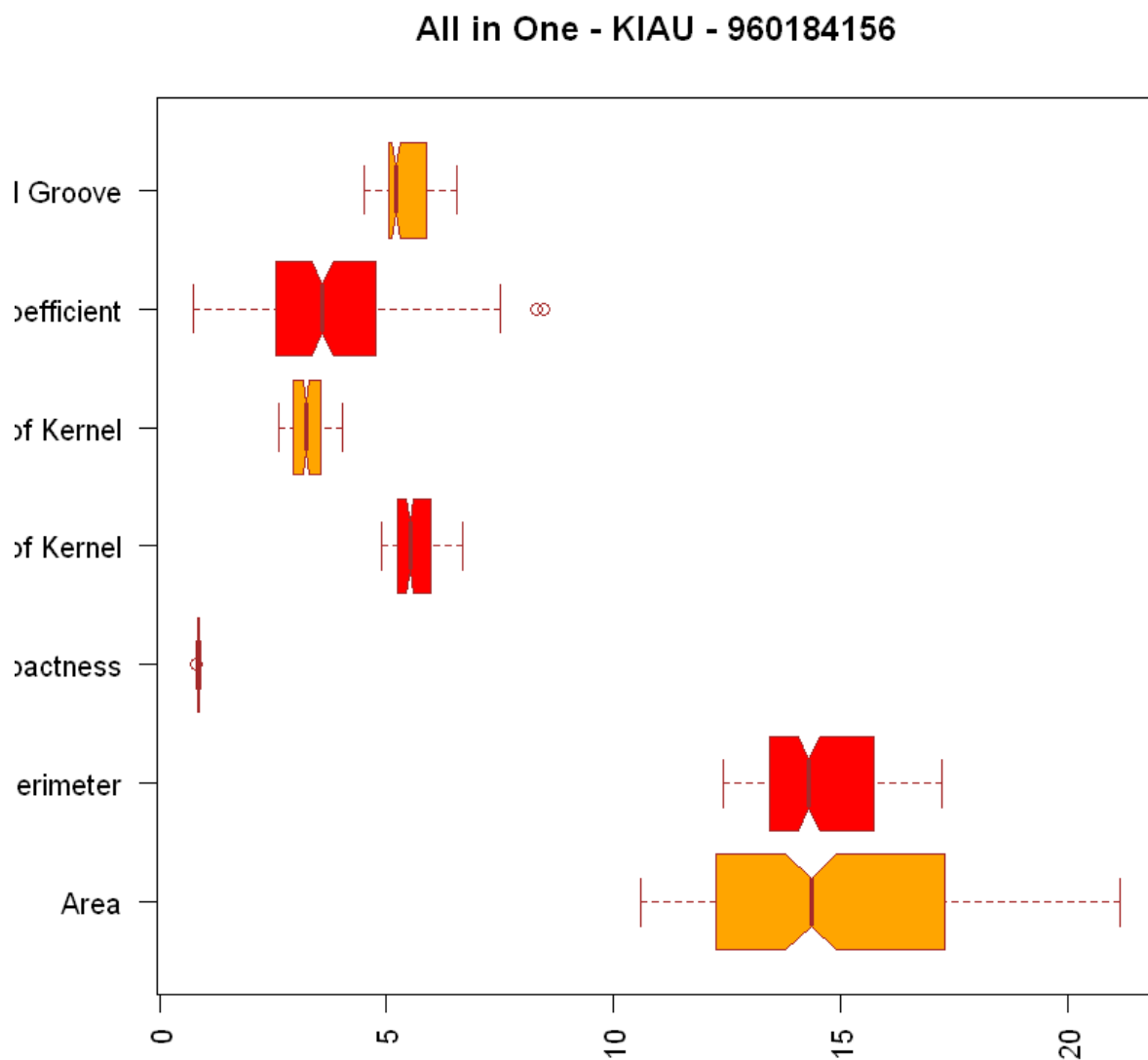
R برای ما امکان شخصی سازی در ترسیم **boxplot** را هم فراهم می کند. طوری که کد زیر منجر به این باکس پلات خواهد شد:

```
boxplot(seeds$Area,  
main = "KIAU - DM - Ali Rezaeinejad",  
xlab = "Parts Per Billion",  
ylab = "Ozone",  
col = "orange",  
border = "brown",  
horizontal = TRUE,  
notch = TRUE)
```



می‌توان با ترکیب صفحات گذشته با یکدیگر و استفاده از تابع **boxplot** با چند ورودی، همه نمودارها را ادغام کرد:

(کد در فایل ارسال شده موجود می‌باشد - خطوط ۱۱۶ تا ۱۲۳)



Variance & Standard Deviation واریانس و انحراف از معیار

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat Seeds I
```

```
In [45]: print ("Variances in Order:")
var(seeds$Area)
var(seeds$Perimeter)
var(seeds$Compactness)
var(seeds$Length.of.Kernel)
var(seeds$Width.of.Kernel)
var(seeds$Asymmetry.Coefficient)
var(seeds$Length.of.Kernel.Groove)
```

```
[1] "Variances in Order:"
```

```
8.46635077694236
```

```
1.70552819548872
```

```
0.00055834932809296
```

```
0.196305245295056
```

```
0.142668201891091
```

```
2.26068404564502
```

```
0.241553080997949
```

```
In [46]: print ("Standard Deviations in Order:")
sd(seeds$Area)
sd(seeds$Perimeter)
sd(seeds$Compactness)
sd(seeds$Length.of.Kernel)
sd(seeds$Width.of.Kernel)
sd(seeds$Asymmetry.Coefficient)
sd(seeds$Length.of.Kernel.Groove)
```

```
[1] "Standard Deviations in Order:"
```

```
2.90969943068736
```

```
1.30595872656402
```

```
0.0236294165838465
```

```
0.443063477726449
```

```
0.377714444906587
```

```
1.50355713082178
```

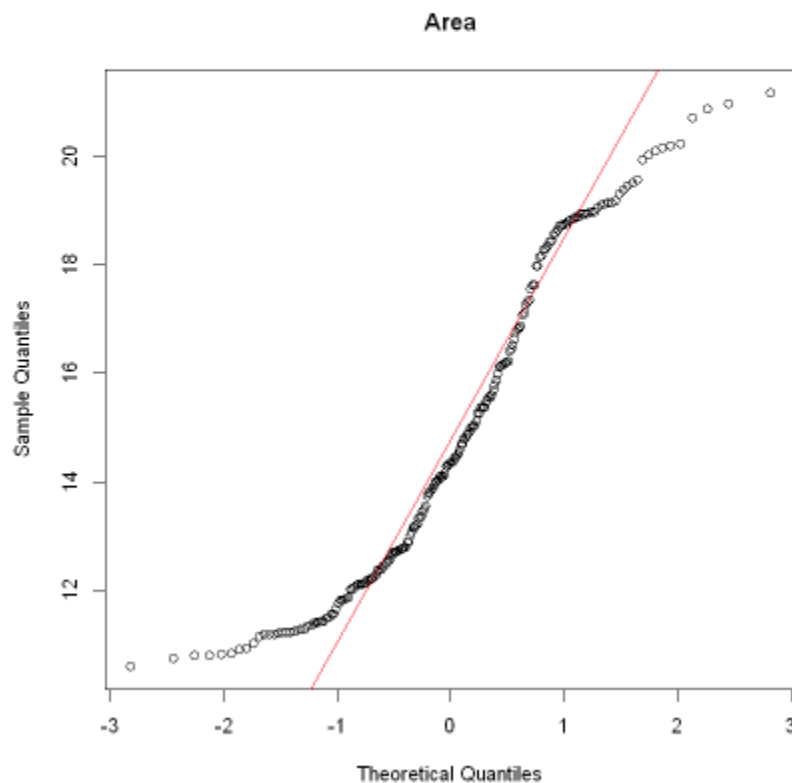
```
0.491480499102405
```

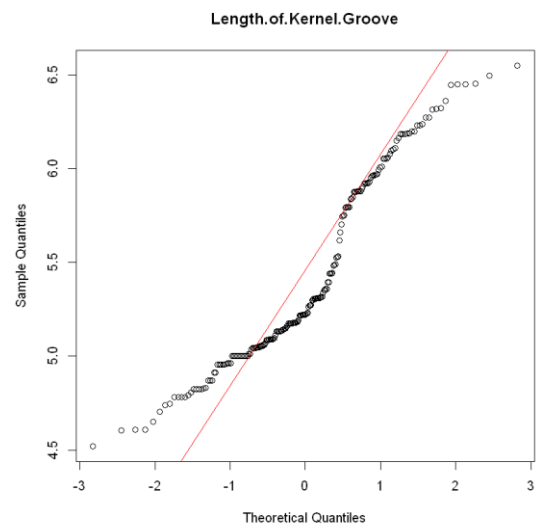
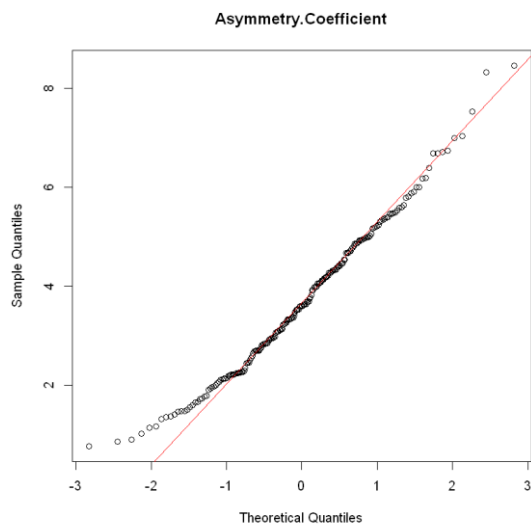
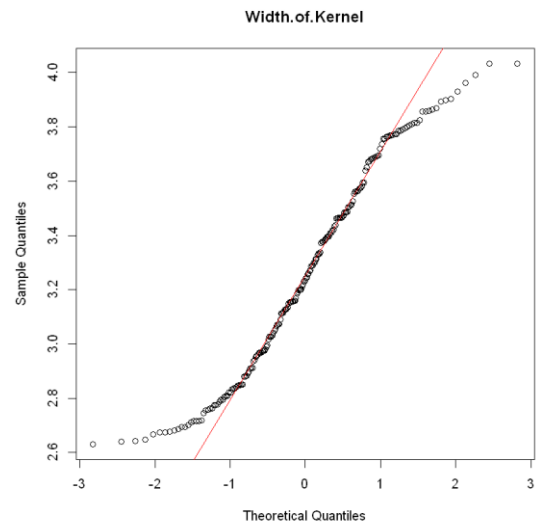
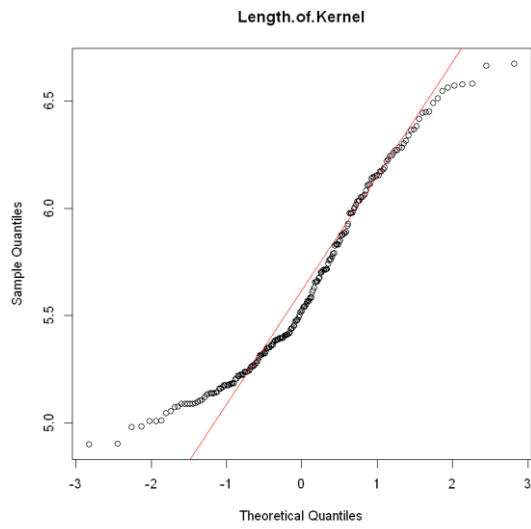
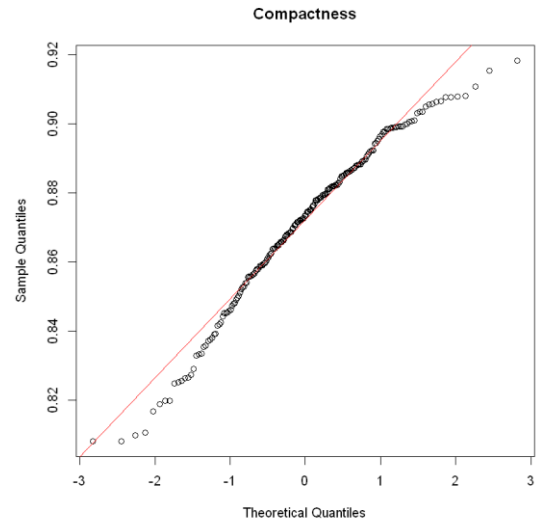
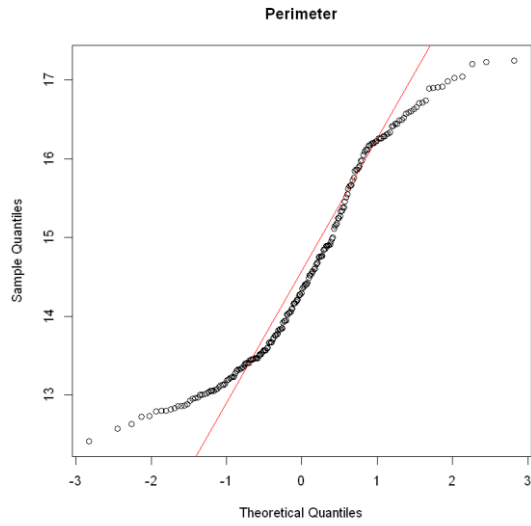
نمایش گرافیکی از توصیف آماری پایه‌ای داده

Q-Q (Quantile-Quantile) Plots

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alterna
```

```
In [11]: qqnorm(seeds$Area, main = "Area", )  
         qqline(seeds$Area,  
               col = "red")
```

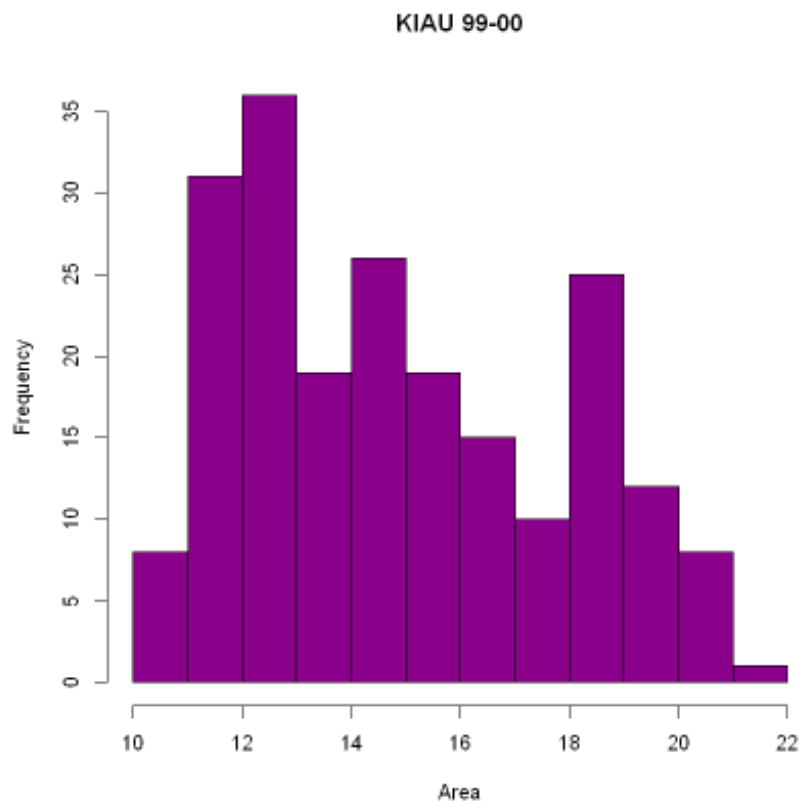


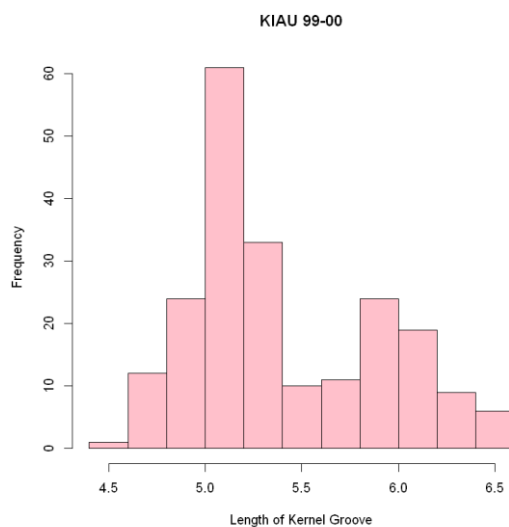
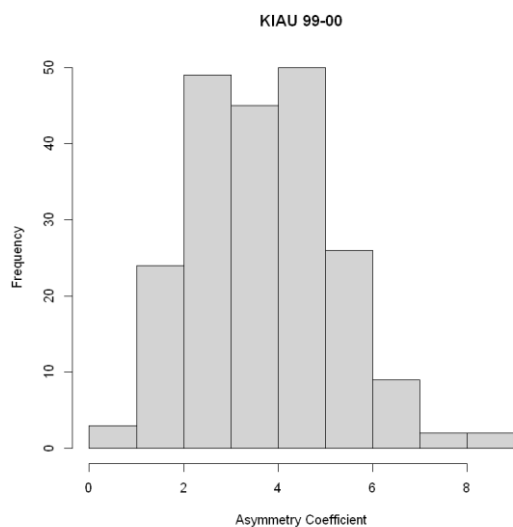
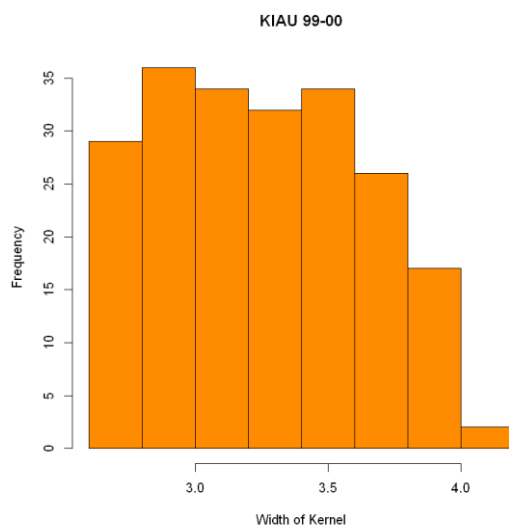
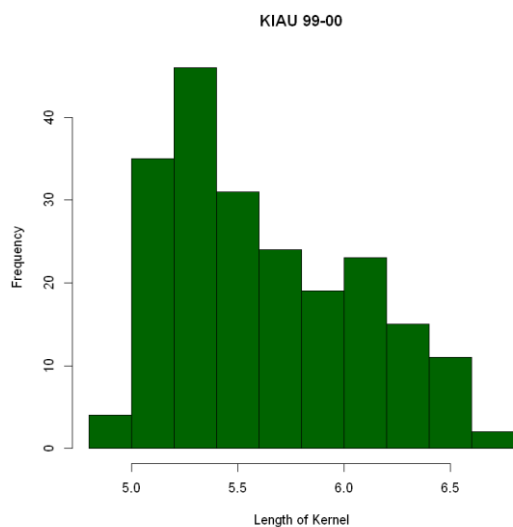
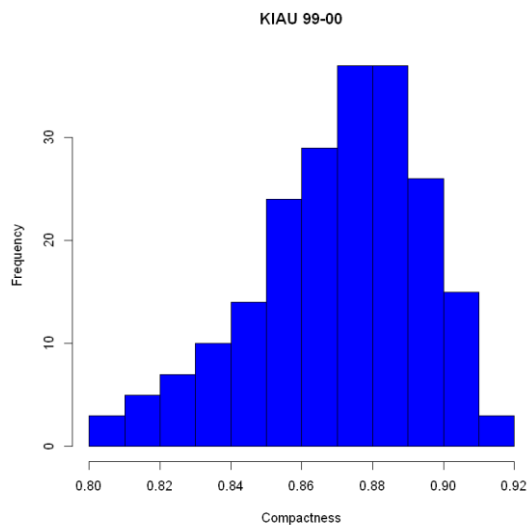
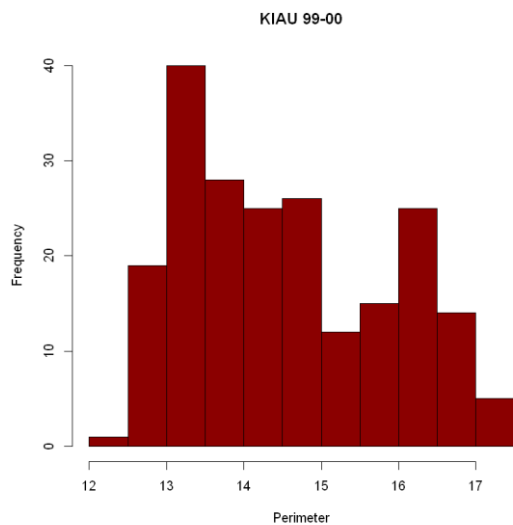


Histogram

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternative")
```

```
In [20]: hist(seeds$Area, main = "KIAU 99-00", xlab="Area", col="darkmagenta")
#hist(seeds$Perimeter, main = "KIAU 99-00", xlab="Perimeter", col="darkmagenta")
#hist(seeds$Compactness, main = "KIAU 99-00", xlab="Compactness", col="darkmagenta")
#hist(seeds$Length.of.Kernel, main = "KIAU 99-00", xlab="Length of Kernel", col="darkmagenta")
#hist(seeds$Width.of.Kernel, main = "KIAU 99-00", xlab="Width of Kernel", col="darkmagenta")
#hist(seeds$Asymmetry.Coefficient, main = "KIAU 99-00", xlab="Asymmetry Coefficient", col="darkmagenta")
#hist(seeds$Length.of.Kernel.Groove, main = "KIAU 99-00", xlab="Length of Kernel Groove", col="darkmagenta")
```

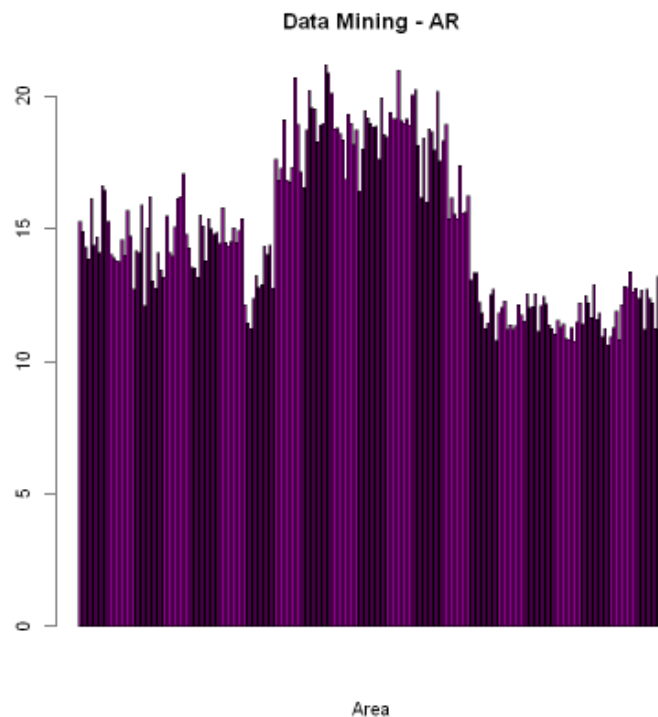


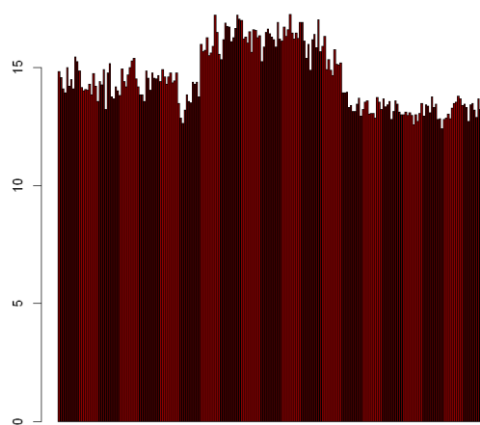


Bar Plot (for all values)

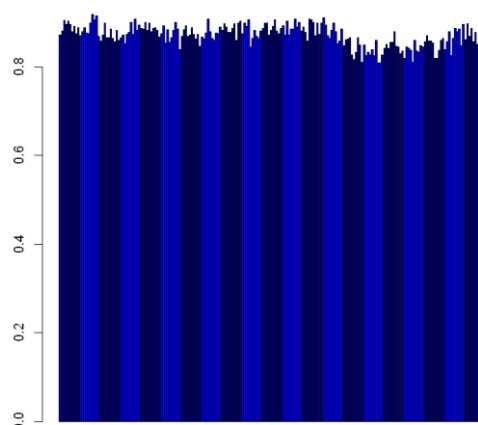
```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternatives\\Wheat
```

```
In [24]: barplot(seeds$Area, main = "Data Mining - AR", xlab="Area", col="darkmagenta")  
#barplot(seeds$Perimeter, xlab="Perimeter", col="darkred")  
#barplot(seeds$Compactness, xlab="Compactness", col="blue")  
#barplot(seeds$Length.of.Kernel, xlab="Length of Kernel", col="darkgreen")  
#barplot(seeds$Width.of.Kernel, xlab="Width of Kernel", col="darkorange")  
#barplot(seeds$Asymmetry.Coefficient, xlab="Asymmetry Coefficient")  
#barplot(seeds$Length.of.Kernel.Groove, xlab="Length of Kernel Groove", col="pi
```

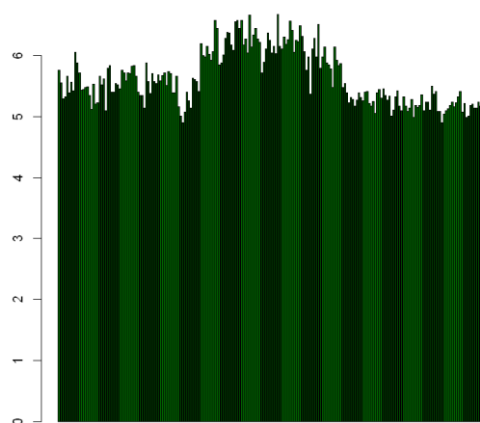




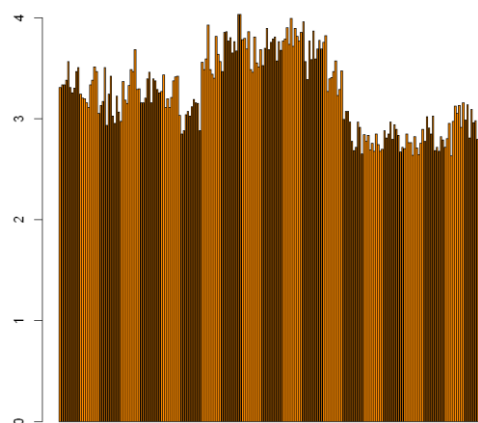
Perimeter



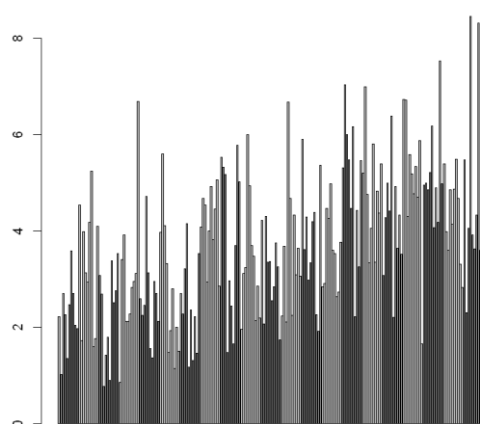
Compactness



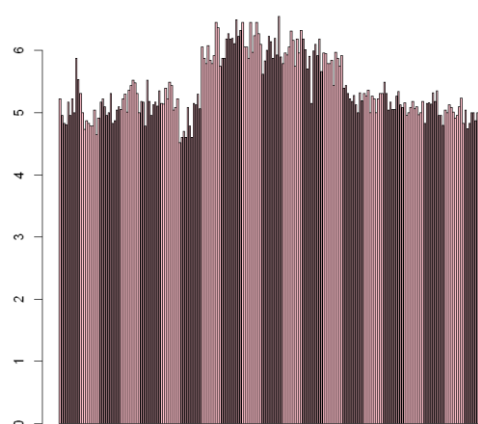
Length of Kernel



Width of Kernel



Asymmetry Coefficient



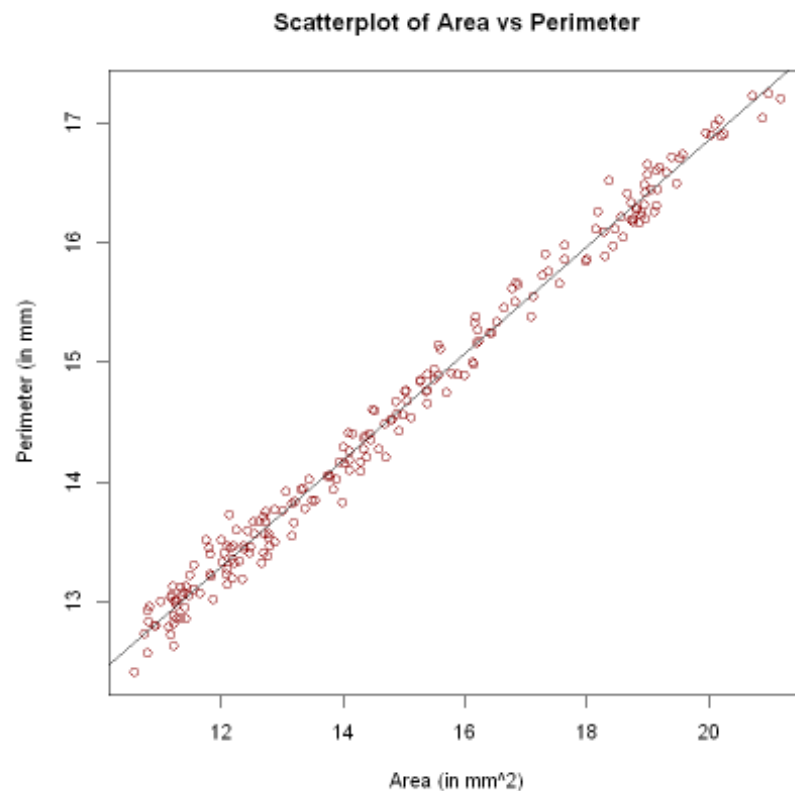
Length of Kernel Groove

Scatter Plot

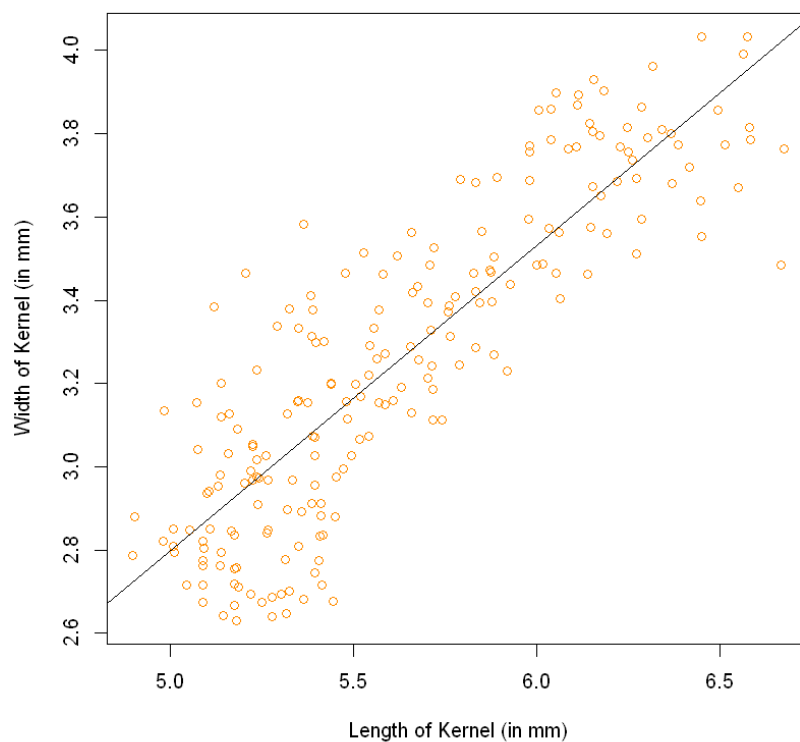
Area – Perimeter (Positive Correlation)

```
In [1]: seeds <- read.csv("D:\\Daneshga\\T8\\Amini\\Datasets\\Iris Alternativ
```

```
In [36]: plot(seeds$Area, seeds$Perimeter,  
             main = "Scatterplot of Area vs Perimeter",  
             xlab = "Area (in mm^2)",  
             ylab = "Perimeter (in mm)", col="darkred")  
abline(lm(seeds$Perimeter ~ seeds$Area))
```



Scatterplot of Length of Kernel vs Width of Kernel



Scatterplot of Length of Kernel vs Length of Kernel Groove

