Data 621 Homework 5: Code Appendix

Jeff Nieman, Scott Karr, James Topor, Armenoush Aslanian-Persico

Contents

This Appendix contains all of the source R code and associated relevant output from our final writeup and our model building efforts for Assignment # 5. The R code is organized to match up to the relevant sections of the Writeup document.

However, we begin here by providing the full ouput of our Evaluation data set predictions as indicated in Part 4 of the final writeup document.

Full Results of Evaluation Data Set Predictions

The full set of Evaluation data set predictions listed in order of their 'INDEX' identifier is provided below. Please note that the estimated probability of a policyholder being involved in a car accident is indicated by the 'TARGET_FLAG_PROB' variable shown in the table.

INDEX	Predicted
3	2
9	4
10	3
18	2
21	1
30	6
31	4
37	1

INDEX	Predicted
39	0
47	1
60	3
62	1
63	4
64	1
68	1
75	3
76	2
83	0
87	4
92	5
98	2
106	1
107	1
113	2
120	4
123	6
125	3
126	6
128	5
129	3
131	4
135	1
141	4
147	3
148	1
151	4
156	4
157	3
174	2
186	1
193	3
195	1
212	1
213	1
217	3
223	4
226	3
228	5
230	4
241	2
243	4
249	1
281	4
288	0
294	2
295	2
300	6
302	5
303	1
308	2

INDEX	Predicted
319	5
320	1
324	3
331	3
343	3
347	2
348	4
350	5
357	1
358	4
360	4
366	4
367	3
368	5
376	2
380	3
388	1
396	5
398	5
403	4
410	2
412	1
420	2
434	2
440	3
450	4
453	3
464	5
465	4
466	5
473	3
476	1
478	2
479	3
493	3
497	3
503	4
504	4
505	2
507	0
513	3
519	3
521	4
522	4
545	3
549	2
551	1
556	7
557	6
559	2
560	3
566	4

INDEX	Predicted
569	5
573	3
578	1
579	5
582	5
596	4
598	2
599	1
602	3
605	2
617	3
619	6
630	3
634	3
643	2
645	1
647	5
649	3
656	4
657	5
658	3
667	3
692	3
693	4
698	0
699	4
700	7
704	3
707	3
708	5
709	3
713	1
714	3
716	2
718	4
722	4
729	7
731	2
733	4
746	3
747	4
748	1
753	1
757	1
763	3
767	5
774	3
776	1
788	1
794	4
799	2
803	4

INDEX	Predicted
806	4
807	4
811	4
816	7
818	3
819	2
831	5
835	5
837	2
841	1
846	1
856	6
861	4
862	1
863	3
865	4
871	3
879	1
880	3
881	3
885	4
887	3
892	1
898	4
900	1
904	2
906	5
910	4
912	4
913	2
919	5
924	1
925	3
930	3
940	2
941	4
946	0
949	5
951	1
962	4
966	2
967	5
971	1
981	3
982	3
983	1
984	1
989	2
990	5
992	3
995	5
996	2

INDEX	Predicted
998	1
1001	3
1007	1
1008	3
1016	3
1022	2
1027	5
1032	1
1033	4
1041	5
1065	1
1074	2
1075	1
1081	1
1094	4
1099	3
1105	3
1123	1
1135	1
1142	2
1155	2
1169	3
1176	4
1178	4
1180	4
1184	1
1185	2
1193	0
1196	1
1199	2
1203	3
1205	3
1207	3
1208	2
1212	1
1213	1
1222	1
1223	2
1226	4
1227	6
1229	1
1230	4
1231	3
1241	1
1243	4
1244	6
1246	4
1248	3
1249	4
1252	3
1261	3
1275	3

INDEX	Predicted
1281	1
1285	4
1288	1
1290	4
1291	1
1304	5
1305	3
1323	5
1342	1
1348	3
1353	4
1363	3
1371	4
1372	1
1378	1
1381	4
1382	4
1393	5
1394	5
1398	7
1404	3
1405	4
1419	1
1421	2
1426	1
1431	1
1435	3
1437	2
1438	2
1442	1
1464	1
1471	3
1473	5
1476	3
1478	2
1479	4
1487	5
1492	4
1496	3
1497	1
1515	3
1519	1
1522	3
1526	4
1537	3
1538	4
1540	2
1543	4
1548	1
1549	1
1556	3
1564	1

INDEX	Predicted
1570	3
1577	2
1585	4
1590	5
1592	2
1594	1
1596	7
1598	5
1603	2
1607	1
1612	6
1627	5
1629	4
1630	3
1640	6
1641	4
1646	4
1662	0
1668	1
1671	0
1672	4
1673	6
1686	4
1688	4
1696	3
1701	5
1707	4
1708	4
1713	4
1715	3
1717	3
1721	3
1724	3
1725	3
1730	4
1731	4
1734	3
1740	4
1748	3
1749	4
1750	6
1763	3
1768	4
1773	2
1777	1
1778	0
1780	3
1782	0
1784	5
1786	4
1787	5
1792	1

INDEX	Predicted
1800	3
1801	3
1803	3
1804	3
1807	2
1818	7
1821	4
1822	4
1828	3
1833	4
1844	4
1847	3
1850	3
1854	4
1858	5
1864	4
1867	1
1876	3
1880	1
1881	2
1891	3
1894	3
1895	5
1901	1
1905	5
1912	6
1918	3
1921	4
1923	3
1924	1
1931	2
1941	4
1950	2
1951	4
1954	5
1961	4
1966	4
1979	4
1982	1
1987	3
1997	3
2004	4
2011	5
2015	4
2025	7
2033	1
2034	5
2035	2
2036	2
2053	3
2059	3
2060	2

INDEX	Predicted
2073	3
2084	3
2089	3
2092	2
2109	5
2129	5
2134	6
2135	5
2148	3
2149	1
2150	2
2165	3
2166	3
2168	5
2170	1
2171	3
2172	3
2176	5
2182	3
2189	2
2191	5
2197	3
2202	0
2203	4
2204	0
2206	5
2218	2
2219	4
2221	1
2226	1
2228	3
2232	4
2236	0
2241	2
2245	5
2251	3
2255	5
2256	4
2259	1
2263	4
2264	4
2267	1
2273	2
2277	5
2287	4
2289	3
2291	1
2296	2
2299	2
2306	3
2314	0
2317	3

INDEX	Predicted
2318	4
2321	4
2324	4
2340	3
2343	2
2349	1
2352	5
2353	2
2365	2
2370	1
2378	3
2390	1
2399	1
2402	1
2403	1
2404	2
2414	5
2422	4
2424	1
2430	4
2435	3
2439	2
2442	5
2445	4
2449	2
2451	1
2461	3
2464	4
2465	4
2472	3
2476	3
2482	3
2487	5
2498	5
2501	4
2504	3
2511	0
2518	5
2521	5
2530	3
2543	5
2545	4
2561	3
2566	4
2572	3
2577	3
2578	2
2580	2
2581	4
2582	5
2584	3
2590	4

INDEX	Predicted
2598	3
2602	1
2605	5
2616	4
2618	4
2619	2
2624	5
2632	1
2640	4
2646	4
2651	1
2660	1
2661	3
2668	1
2670	1
2680	3
2681	3
2689	1
2694	6
2695	1
2696	3
2702	3
2704	4
2708	6
2709	2
2714	4
2716	1
2723	1
2725	3
2738	0
2750	4
2756	1
2758	4
2766	3
2767	0
2771	2
2775	4
2776	3
2779	2
2780	3
2781	1
2782	4
2783	4
2796	4
2798	5
2800	3
2803	7
2806	5
2813	3
2818	2
2821	5
2825	4

INDEX	Predicted
2829	2
2830	3
2833	4
2839	1
2843	4
2846	5
2847	3
2848	6
2856	4
2863	1
2867	5
2869	4
2873	4
2874	4
2875	3
2880	2
2886	3
2887	4
2888	3
2889	3
2890	2
2892	2
2901	3
2902	2
2905	2
2917	2
2922	3
2924	3
2930	1
2931	5
2946	2
2955	4
2962	3
2964	1
2965	1
2967	6
2970	6
2973	7
2974	2
2976	1
2977	2
2978	1
2986	4
2988	5
2989	1
2995	4
3005	5
3011	5
3013	3
3019	4
3021	2
3022	5

INDEX	Predicted
3029	1
3037	4
3042	3
3043	4
3049	4
3050	6
3053	1
3058	1
3062	1
3063	4
3065	2
3080	3
3088	1
3093	1
3096	5
3101	6
3103	2
3107	6
3109	4
3111	5
3113	5
3116	5
3132	3
3141	5
3153	1
3154	1
3160	1
3167	1
3170	3
3173	2
3174	4
3177	5
3179	4
3184	1
3190	3
3193	4
3199	4
3201	1
3202	3
3203	3
3206	3
3209	3
3210	3
3217	4
3220	4
3228	3
3232	3
3239	1
3243	4
3245	3
3246	4
3251	5

INDEX	Predicted
3253	5
3257	4
3260	1
3261	4
3263	5
3278	4
3281	3
3283	5
3290	2
3297	4
3304	4
3305	3
3307	5
3308	3
3313	3
3314	3
3317	0
3348	1
3350	2
3359	4
3367	4
3376	3
3378	3
3384	1
3386	3
3387	4
3388	3
3390	5
3391	1
3396	0
3398	4
3404	5
3406	1
3407	4
3414	5
3419	3
3423	3
3427	3
3432	0
3434	3
3438	1
3442	0
3443	1
3448	0
3456	3
3464	6
3470	1
3475	4
3477	3
3490	3
3493	4
3502	3

INDEX	Predicted
3508	4
3516	2
3517	5
3525	3
3532	4
3535	4
3536	5
3540	3
3547	3
3550	3
3557	2
3562	1
3563	3
3564	1
3570	3
3573	3
3577	3
3579	3
3581	1
3587	4
3602	3
3609	4
3612	4
3621	3
3642	1
3647	1
3649	3
3654	2
3660	5
3665	4
3669	4
3673	4
3675	1
3678	4
3680	4
3686	6
3693	5
3710	3
3713	4
3718	5
3725	3
3726	3
3747	2
3753	1
3754	5
3760	5
3763	3
3765	4
3769	5
3771	4
3784	3
3787	5

INDEX	Predicted
3794	1
3796	4
3798	5
3809	4
3812	5
3819	1
3828	4
3831	5
3833	3
3837	5
3839	1
3843	2
3846	4
3854	5
3861	1
3864	4
3868	3
3869	5
3870	3
3883	3
3886	3
3889	4
3894	1
3907	3
3910	4
3913	1
3914	3
3921	6
3923	1
3929	2
3931	2
3932	5
3937	0
3943	3
3956	5
3957	3
3961	6
3971	4
4004	1
4005	3
4006	4
4011	3
4013	5
4014	7
4016	0
4017	5
4020	3
4022	4
4026	1
4032	1
4043	2
4045	3

INDEX	Predicted
4048	5
4051	5
4052	4
4056	3
4059	2
4069	5
4074	4
4076	3
4077	2
4079	1
4081	4
4088	2
4105	3
4125	4
4134	3
4139	2
4146	2
4149	4
4151	0
4155	3
4157	2
4168	5
4170	2
4174	3
4179	5
4185	4
4199	1
4205	1
4208	3
4211	2
4212	0
4215	3
4217	3
4219	2
4226	4
4227	3
4229	2
4231	2
4233	2
4237	3
4243	5
4248	4
4255	5
4262	3
4266	1
4268	1
4270	2
4273	1
4276	4
4277	4
4279	2
4299	4

INDEX	Predicted
4313	1
4322	4
4324	1
4328	3
4331	3
4335	2
4337	4
4338	1
4343	2
4347	2
4355	4
4357	3
4359	6
4362	2
4368	3
4374	5
4375	5
4378	4
4381	2
4387	5
4400	2
4423	4
4424	3
4428	4
4433	7
4436	2
4437	1
4439	6
4449	4
4456	4
4463	6
4467	2
4468	2
4469	3
4472	4
4473	4
4476	3
4500	2
4509	4
4513	2
4521	1
4527	3
4530	2
4532	3
4533	3
4535	3
4536	5
4542	4
4551	3
4554	4
4555	2
4564	1

INDEX	Predicted
4572	2
4573	7
4577	2
4579	5
4583	4
4584	6
4596	3
4599	4
4607	4
4609	0
4610	1
4616	2
4617	3
4633	4
4638	4
4641	3
4653	7
4655	4
4659	1
4669	1
4678	1
4685	5
4686	4
4691	1
4695	5
4698	4
4700	5
4711	3
4722	4
4727	3
4756	6
4762	1
4763	3
4766	4
4770	1
4784	3
4791	3
4795	5
4799	1
4802	4
4805	4
4814	5
4816	1
4817	4
4822	3
4827	3
4833	5
4836	1
4842	3
4844	3
4845	3
4849	4

INDEX	Predicted
4850	1
4860	4
4863	3
4871	4
4878	4
4881	3
4888	3
4900	7
4906	4
4909	1
4916	5
4918	6
4926	3
4928	3
4941	3
4946	4
4949	1
4956	3
4966	4
4969	3
4973	4
4978	5
4982	3
4985	5
4991	4
4998	4
5000	3
5004	3
5005	2
5011	5
5016	0
5018	3
5034	4
5038	1
5042	7
5046	3
5051	0
5054	1
5057	4
5062	4
5063	4
5065	3
5066	2
5076	3
5089	3
5092	2
5093	4
5094	4
5098	4
5102	3
5112	5
5117	3

INDEX	Predicted
5127	3
5130	3
5131	2
5132	4
5135	1
5136	3
5147	5
5157	4
5160	2
5165	1
5166	1
5172	3
5173	3
5179	1
5184	5
5187	2
5191	3
5193	2
5194	1
5199	2
5212	1
5213	3
5224	4
5226	4
5239	5
5252	4
5264	0
5266	2
5271	5
5273	4
5276	2
5278	4
5281	2
5283	4
5291	0
5294	7
5296	4
5297	1
5313	4
5314	3
5321	3
5325	3
5326	1
5328	3
5334	3
5338	5
5344	2
5348	2
5352	2
5353	4
5354	1
5361	0

INDEX	Predicted
5364	3
5365	4
5367	2
5379	5
5382	3
5386	5
5395	4
5410	5
5411	3
5416	5
5424	4
5426	3
5428	2
5430	5
5433	1
5437	3
5440	4
5442	5
5445	3
5449	4
5452	4
5460	1
5461	2
5465	2
5467	4
5471	5
5474	1
5475	3
5480	1
5481	4
5484	2
5494	6
5495	2
5497	1
5499	3
5507	1
5510	3
5515	1
5516	2
5517	1
5524	4
5530	6
5534	3
5543	1
5545	3
5558	4
5562	2
5573	7
5581	4
5583	5
5587	3
5589	2

INDEX	Predicted
5591	5
5596	3
5606	4
5608	5
5611	3
5612	5
5614	4
5620	4
5623	5
5624	2
5626	6
5633	1
5635	3
5640	5
5643	3
5644	6
5653	5
5663	4
5664	5
5667	3
5671	1
5673	4
5676	2
5678	2
5698	2
5700	6
5705	4
5706	5
5711	2
5712	5
5716	3
5719	3
5725	2
5728	7
5734	1
5735	4
5743	3
5754	2
5755	3
5756	5
5766	3
5770	4
5774	1
5775	2
5776	4
5778	6
5786	3
5787	4
5791	5
5794	3
5803	3
5804	4

INDEX	Predicted
5808	1
5810	4
5813	4
5828	3
5839	6
5842	4
5843	3
5844	4
5847	3
5851	3
5854	3
5857	1
5866	2
5874	1
5886	4
5895	1
5897	3
5898	3
5900	2
5902	1
5908	0
5909	3
5912	6
5913	1
5917	5
5918	3
5921	3
5931	2
5942	4
5943	3
5950	4
5954	1
5983	3
5995	3
6002	3
6005	4
6009	7
6011	3
6012	5
6019	2
6021	3
6029	5
6036	4
6037	3
6038	1
6043	1
6045	4
6047	1
6048	3
6061	4
6063	3
6064	5

INDEX	Predicted
6068	4
6069	1
6070	4
6071	5
6074	4
6079	4
6082	3
6088	4
6094	3
6095	2
6098	3
6102	4
6105	2
6113	5
6116	3
6120	2
6121	3
6126	2
6144	4
6145	4
6153	3
6156	4
6159	4
6162	1
6184	3
6188	3
6189	3
6191	2
6211	1
6216	4
6218	1
6222	1
6235	3
6245	1
6248	4
6253	4
6256	0
6257	4
6259	1
6266	5
6268	4
6275	1
6280	3
6283	3
6288	5
6289	2
6301	6
6308	4
6314	4
6315	0
6316	4
6317	3

INDEX	Predicted
6318	3
6323	3
6329	3
6336	1
6341	5
6348	3
6349	5
6365	1
6372	3
6376	3
6378	1
6379	5
6382	1
6383	3
6389	5
6390	2
6392	4
6394	1
6402	1
6404	4
6405	3
6406	1
6409	4
6410	5
6411	4
6421	5
6428	7
6429	3
6432	4
6436	1
6437	4
6438	0
6445	3
6447	4
6450	5
6462	5
6467	4
6478	4
6484	5
6492	5
6497	6
6504	3
6505	1
6513	3
6525	5
6526	5
6528	3
6540	1
6542	1
6544	5
6548	4
6552	3

INDEX	Predicted
6558	4
6567	2
6569	4
6572	6
6577	6
6581	3
6588	3
6591	1
6594	3
6600	6
6602	4
6604	3
6605	4
6614	3
6616	2
6621	1
6640	4
6641	3
6643	2
6644	1
6649	2
6650	4
6655	6
6661	2
6672	5
6677	2
6688	2
6689	5
6691	3
6692	5
6694	3
6702	2
6714	1
6716	5
6724	4
6725	2
6730	4
6735	3
6738	3
6739	3
6743	4
6747	3
6750	7
6751	3
6753	4
6754	3
6755	2
6762	4
6764	4
6772	2
6774	1
6787	2

INDEX	Predicted
6789	3
6793	4
6798	1
6799	0
6800	4
6802	2
6808	2
6809	4
6812	4
6814	3
6816	3
6822	1
6829	4
6834	3
6836	4
6839	3
6840	2
6843	1
6846	2
6848	2
6852	2
6856	6
6860	5
6866	4
6870	4
6878	1
6880	5
6885	6
6897	4
6902	3
6904	4
6907	4
6909	1
6914	2
6915	5
6922	1
6924	2
6933	3
6934	5
6941	4
6957	3
6960	1
6969	0
6975	3
6980	2
6983	2
6987	2
6994	3
6997	5
7002	4
7010	2
7015	3

INDEX	Predicted
7019	3
7022	4
7025	3
7029	4
7031	3
7037	2
7038	4
7043	2
7049	4
7052	3
7053	4
7056	3
7057	4
7080	1
7086	2
7087	4
7105	4
7108	2
7121	1
7122	2
7125	2
7132	4
7134	4
7151	3
7152	3
7157	3
7159	1
7166	1
7167	1
7177	3
7179	2
7181	1
7183	5
7186	4
7193	5
7205	3
7207	2
7209	4
7216	1
7232	1
7235	2
7238	3
7240	2
7243	1
7252	2
7269	3
7275	4
7281	3
7283	2
7287	7
7289	6
7291	4

INDEX	Predicted
7294	1
7304	5
7308	4
7313	3
7319	1
7325	2
7326	3
7330	5
7332	5
7337	1
7341	1
7346	3
7353	1
7354	4
7361	1
7366	1
7368	5
7372	4
7375	3
7377	6
7380	1
7382	4
7385	3
7392	2
7395	5
7397	3
7403	3
7406	3
7409	0
7410	4
7412	4
7419	3
7425	3
7435	7
7438	4
7440	4
7447	3
7449	2
7456	3
7464	4
7478	3
7480	3
7481	5
7483	3
7484	2
7491	4
7494	2
7501	2
7503	4
7509	3
7517	0
7518	6

INDEX	Predicted
7519	2
7521	5
7522	3
7536	4
7539	1
7547	7
7549	3
7552	3
7554	2
7556	3
7564	4
7566	1
7570	4
7571	3
7572	4
7575	2
7586	3
7589	5
7590	3
7597	4
7602	5
7604	4
7605	3
7612	2
7615	4
7617	3
7624	4
7632	4
7639	4
7642	3
7643	3
7649	4
7650	2
7653	4
7654	4
7657	5
7662	1
7669	4
7671	3
7675	1
7678	4
7682	4
7688	2
7689	1
7690	5
7692	3
7699	1
7705	3
7712	3
7726	4
7728	2
7735	4

INDEX	Predicted
7737	3
7739	4
7743	1
7744	6
7746	2
7749	1
7750	4
7752	4
7755	5
7756	4
7762	4
7764	1
7769	3
7770	2
7776	5
7778	5
7784	3
7786	3
7789	3
7793	4
7794	2
7804	3
7811	5
7813	3
7815	3
7817	1
7818	4
7821	6
7825	3
7830	5
7832	4
7835	2
7839	7
7842	5
7849	4
7856	3
7857	3
7863	5
7866	4
7871	3
7875	2
7882	2
7887	1
7888	3
7891	2
7895	4
7901	4
7906	3
7908	1
7917	2
7924	3
7948	4

INDEX	Predicted
7950	6
7955	3
7957	1
7959	2
7967	5
7969	3
7971	5
7974	3
7976	3
7986	6
7987	1
7993	4
7996	4
7998	4
8018	3
8019	3
8027	2
8036	2
8040	5
8044	4
8050	3
8052	0
8054	0
8057	4
8058	4
8059	3
8066	3
8070	6
8072	7
8078	3
8079	3
8080	4
8081	4
8088	1
8091	3
8094	3
8095	3
8099	4
8101	5
8102	5
8116	4
8125	4
8134	1
8139	1
8141	5
8147	3
8158	6
8160	3
8165	3
8187	3
8205	1
8209	3

INDEX	Predicted
8211	4
8232	4
8236	5
8237	5
8238	6
8245	4
8256	3
8268	3
8269	3
8270	5
8286	3
8289	4
8301	3
8305	1
8310	3
8312	1
8318	5
8321	4
8328	1
8331	4
8334	3
8344	4
8345	2
8352	4
8358	4
8359	2
8360	3
8365	3
8366	3
8369	5
8373	5
8378	4
8392	3
8397	3
8399	2
8400	3
8405	3
8406	5
8410	1
8413	4
8414	1
8416	5
8426	2
8434	5
8439	1
8440	1
8475	4
8480	5
8497	2
8499	1
8500	4
8501	3

INDEX	Predicted
8502	3
8518	6
8520	4
8523	5
8525	2
8532	3
8535	2
8543	2
8554	1
8560	5
8561	4
8563	1
8566	1
8570	4
8572	3
8582	1
8583	2
8587	2
8592	2
8593	1
8607	1
8609	5
8610	4
8614	4
8616	5
8622	4
8623	4
8624	2
8633	6
8641	6
8644	6
8649	5
8653	3
8657	7
8658	3
8663	4
8672	2
8680	1
8684	3
8687	1
8688	4
8690	4
8712	0
8717	4
8730	5
8739	4
8744	5
8747	5
8748	4
8751	3
8758	4
8761	2

INDEX	Predicted
8763	1
8764	4
8765	7
8773	4
8780	3
8781	3
8782	4
8785	0
8786	2
8797	1
8799	4
8807	1
8816	3
8817	3
8826	5
8833	3
8834	3
8835	2
8840	5
8843	3
8849	3
8855	4
8861	2
8862	4
8865	3
8868	5
8870	5
8880	5
8885	2
8894	2
8895	4
8899	4
8912	1
8922	3
8924	3
8928	4
8932	1
8943	5
8945	6
8946	5
8954	2
8958	4
8960	2
8965	1
8966	5
8967	2
8969	3
8980	4
8984	4
8985	3
8988	3
8989	5

INDEX	Predicted
8995	1
9004	3
9010	2
9012	1
9018	4
9036	1
9037	2
9040	2
9041	6
9044	6
9045	3
9047	2
9049	2
9061	2
9062	3
9076	4
9079	3
9081	4
9082	4
9089	1
9092	3
9094	1
9115	1
9117	4
9118	3
9120	1
9124	3
9128	1
9135	1
9136	3
9138	4
9157	4
9176	2
9183	2
9187	3
9188	2
9190	4
9197	4
9200	3
9201	1
9203	1
9212	4
9213	2
9214	4
9217	3
9219	4
9220	5
9221	6
9237	2
9240	5
9241	2
9248	3

INDEX	Predicted
9253	7
9259	4
9267	1
9271	3
9273	1
9285	6
9290	4
9291	3
9293	1
9294	4
9301	3
9302	4
9312	3
9316	4
9319	1
9328	5
9331	4
9338	1
9350	3
9356	4
9359	2
9362	4
9364	3
9370	4
9380	1
9386	5
9394	2
9407	4
9411	3
9422	4
9423	1
9429	1
9433	1
9439	1
9451	4
9452	4
9453	2
9460	5
9465	6
9470	4
9476	4
9485	4
9486	1
9488	3
9507	6
9508	2
9517	7
9521	6
9528	3
9532	3
9536	3
9540	5

INDEX	Predicted
9542	5
9546	3
9548	4
9549	6
9554	5
9555	5
9558	1
9573	1
9575	5
9584	3
9586	4
9588	5
9591	1
9592	4
9597	7
9600	4
9603	4
9605	3
9614	1
9616	5
9622	4
9624	4
9629	1
9633	4
9640	4
9644	2
9645	1
9646	3
9648	3
9649	1
9660	4
9664	3
9675	1
9679	3
9680	3
9682	1
9697	3
9701	3
9704	3
9705	1
9707	4
9714	1
9718	2
9722	5
9739	3
9747	6
9751	2
9757	2
9759	3
9760	4
9764	3
9776	1

INDEX	Predicted
9778	3
9786	1
9803	4
9804	5
9815	5
9824	2
9825	0
9826	3
9827	4
9833	4
9835	1
9860	4
9865	3
9871	3
9874	2
9880	2
9882	1
9885	3
9888	5
9892	2
9893	5
9896	1
9902	3
9906	4
9910	5
9914	2
9918	1
9920	1
9926	4
9931	6
9935	4
9945	4
9953	1
9957	3
9963	5
9972	4
9976	4
9979	3
9980	1
9982	1
9991	4
10000	6
10003	4
10005	2
10014	3
10032	4
10034	3
10041	3
10042	4
10044	5
10045	1
10054	4

INDEX	Predicted
10061	6
10062	1
10073	1
10081	0
10084	1
10086	0
10093	1
10101	5
10105	4
10110	4
10113	3
10115	3
10119	4
10121	3
10124	1
10126	6
10127	4
10145	3
10147	1
10148	3
10162	2
10163	3
10166	2
10172	4
10173	3
10175	2
10180	3
10186	1
10192	3
10199	3
10209	0
10210	7
10214	3
10215	0
10216	1
10232	3
10239	5
10249	5
10253	5
10255	1
10262	3
10264	3
10266	1
10268	1
10271	3
10272	5
10276	1
10277	2
10279	3
10281	2
10285	2
10294	1

INDEX	Predicted
10300	3
10304	3
10307	3
10309	5
10310	1
10312	1
10321	3
10332	0
10336	4
10368	3
10369	6
10375	1
10376	2
10379	4
10380	3
10383	2
10385	5
10387	4
10397	3
10412	3
10413	1
10418	3
10420	3
10426	4
10427	3
10428	3
10430	1
10435	2
10436	1
10446	6
10448	3
10449	5
10463	2
10469	2
10470	4
10471	4
10473	5
10476	3
10482	3
10500	3
10511	5
10512	4
10514	4
10515	4
10526	1
10546	2
10549	4
10553	1
10558	2
10575	1
10581	3
10583	1

INDEX	Predicted
10584	2
10585	1
10610	2
10611	1
10616	5
10618	3
10628	1
10632	2
10642	4
10648	4
10649	4
10650	3
10654	2
10656	5
10661	5
10663	1
10672	3
10678	5
10685	4
10690	5
10702	4
10706	4
10708	3
10716	3
10717	5
10720	5
10729	1
10730	5
10745	2
10753	2
10754	2
10762	2
10766	1
10776	1
10783	3
10789	3
10790	6
10797	0
10807	3
10810	1
10817	3
10820	3
10822	3
10828	5
10829	3
10830	2
10831	6
10841	6
10847	3
10856	0
10860	0
10861	5

INDEX	Predicted
10863	3
10875	3
10884	4
10895	2
10897	3
10898	3
10903	0
10908	1
10924	3
10926	2
10927	3
10928	3
10933	0
10939	6
10942	4
10945	3
10949	4
10950	3
10958	6
10963	4
10967	3
10971	1
10972	2
10974	4
10976	5
10980	3
10991	2
10995	4
11014	5
11017	5
11019	5
11022	2
11030	4
11031	3
11041	1
11042	4
11044	4
11047	5
11048	3
11049	3
11052	3
11058	1
11069	3
11070	3
11073	5
11074	1
11078	1
11079	1
11085	1
11088	3
11106	1
11110	6

INDEX	Predicted
11114	3
11118	3
11129	5
11130	4
11131	4
11133	3
11138	3
11143	4
11146	4
11153	4
11162	3
11170	7
11171	1
11201	3
11216	5
11219	3
11222	4
11234	1
11238	4
11244	3
11246	2
11248	2
11250	1
11256	5
11259	2
11263	0
11264	3
11270	2
11274	1
11281	3
11285	1
11300	2
11305	3
11317	3
11319	3
11330	1
11334	5
11335	6
11336	3
11356	4
11358	3
11360	2
11364	1
11373	6
11379	3
11382	4
11383	2
11385	4
11387	3
11391	3
11397	3
11404	2

INDEX	Predicted
11405	0
11409	4
11419	3
11430	4
11434	6
11436	2
11440	4
11443	3
11449	2
11452	4
11453	1
11456	5
11457	1
11459	4
11471	0
11476	3
11479	1
11481	4
11485	3
11486	3
11487	3
11488	2
11498	2
11506	3
11511	5
11515	2
11518	1
11521	2
11523	3
11524	1
11525	4
11528	1
11530	3
11531	3
11533	3
11535	4
11537	4
11538	2
11541	2
11548	4
11552	3
11558	1
11560	1
11566	1
11572	1
11573	4
11582	4
11586	2
11590	4
11591	3
11601	4
11611	4

INDEX	Predicted
11617	1
11619	4
11624	4
11626	6
11644	2
11652	1
11656	2
11658	3
11659	5
11663	3
11665	4
11683	4
11685	1
11691	3
11694	4
11698	1
11700	4
11703	3
11705	3
11710	3
11711	3
11714	2
11731	5
11732	0
11742	3
11744	4
11745	3
11749	1
11756	4
11761	1
11762	4
11766	4
11767	8
11769	3
11770	4
11771	4
11777	3
11778	5
11779	1
11788	2
11790	5
11794	4
11801	4
11807	1
11812	4
11817	1
11818	3
11825	1
11828	4
11833	4
11837	5
11838	0

INDEX	Predicted
11842	2
11853	4
11857	4
11858	2
11860	5
11867	4
11868	4
11871	4
11875	3
11881	6
11890	1
11892	4
11894	3
11896	5
11903	4
11905	3
11907	3
11909	7
11911	3
11915	2
11918	3
11920	5
11923	4
11924	3
11926	2
11931	4
11933	4
11940	5
11951	3
11953	3
11973	1
11984	2
11985	2
11991	3
12002	3
12006	3
12008	4
12013	1
12015	3
12016	4
12023	2
12029	2
12036	1
12038	4
12041	1
12049	4
12050	1
12054	4
12060	3
12062	5
12065	3
12079	2

INDEX	Predicted
12083	6
12090	5
12091	5
12094	4
12099	4
12101	3
12110	3
12116	6
12122	3
12127	7
12133	4
12142	1
12147	4
12156	3
12157	5
12158	6
12161	4
12163	4
12166	3
12170	0
12174	5
12183	1
12188	1
12189	4
12192	3
12201	3
12204	0
12207	1
12208	2
12209	3
12210	7
12217	3
12227	1
12231	3
12232	4
12239	5
12240	2
12251	3
12256	5
12261	0
12263	2
12266	1
12267	3
12268	4
12279	5
12280	3
12283	3
12284	0
12285	6
12286	5
12292	5
12295	2

INDEX	Predicted
12301	4
12314	2
12315	0
12318	1
12332	5
12334	2
12337	4
12338	5
12349	4
12350	3
12359	6
12360	5
12373	5
12374	2
12380	4
12382	3
12383	3
12390	5
12398	5
12405	4
12407	3
12410	6
12418	7
12421	5
12422	3
12439	2
12444	4
12463	5
12465	1
12470	4
12471	4
12480	4
12482	4
12484	3
12487	4
12491	4
12503	4
12507	1
12526	3
12533	3
12540	1
12543	3
12552	1
12555	7
12556	4
12570	4
12579	3
12588	3
12600	1
12615	5
12624	4
12629	2

INDEX	Predicted
12634	3
12638	1
12646	3
12650	0
12665	3
12674	3
12676	0
12678	3
12685	3
12690	3
12698	4
12702	5
12704	3
12705	3
12710	3
12715	5
12720	4
12734	3
12744	4
12747	2
12757	5
12758	2
12766	1
12782	3
12787	1
12799	2
12804	5
12809	3
12813	3
12816	3
12821	4
12826	2
12831	4
12832	4
12833	3
12835	4
12842	2
12844	4
12847	1
12852	3
12856	4
12857	3
12858	6
12861	3
12869	4
12876	5
12877	2
12879	3
12882	5
12883	4
12887	2
12889	4

INDEX	Predicted
12891	5
12894	3
12895	1
12899	2
12905	7
12913	2
12916	1
12917	3
12925	3
12934	5
12939	5
12943	2
12950	5
12961	0
12963	3
12973	1
12979	4
12980	1
12981	1
12982	2
12992	2
12994	1
12999	3
13002	4
13004	1
13010	3
13013	3
13015	5
13019	6
13030	2
13031	5
13036	4
13037	5
13042	1
13054	3
13060	2
13072	3
13073	3
13079	5
13081	2
13086	3
13087	4
13090	2
13098	3
13100	3
13105	1
13106	3
13107	4
13113	7
13115	4
13117	1
13118	3

INDEX	Predicted
13121	1
13137	2
13146	1
13150	5
13151	4
13152	3
13156	3
13165	6
13169	4
13178	4
13180	4
13183	4
13184	0
13188	3
13191	4
13196	1
13203	4
13206	3
13211	5
13219	1
13223	6
13226	4
13228	2
13230	5
13240	5
13249	5
13250	2
13256	4
13261	2
13263	2
13268	5
13275	1
13277	5
13283	4
13284	1
13285	3
13286	4
13287	2
13290	3
13291	1
13294	5
13295	3
13303	3
13306	1
13311	5
13322	1
13331	1
13337	5
13344	1
13362	5
13364	2
13366	2

INDEX	Predicted
13368	3
13370	2
13377	5
13378	2
13388	0
13392	5
13398	5
13403	5
13404	5
13409	4
13416	1
13422	1
13427	4
13433	5
13438	3
13441	6
13449	5
13450	4
13453	2
13460	3
13461	5
13465	4
13468	1
13481	4
13485	2
13487	5
13490	4
13493	3
13497	3
13508	1
13516	0
13525	3
13533	4
13535	1
13538	4
13545	4
13566	3
13581	3
13584	2
13588	3
13596	3
13600	7
13604	4
13608	3
13611	3
13612	1
13615	4
13616	4
13618	4
13625	1
13628	3
13629	0

INDEX	Predicted
13630	5
13633	4
13637	4
13640	2
13641	3
13651	5
13674	4
13684	4
13690	2
13707	4
13709	5
13710	3
13713	5
13724	1
13725	1
13731	2
13736	2
13740	1
13745	6
13748	1
13751	1
13758	2
13762	0
13764	4
13765	2
13769	2
13770	3
13774	5
13787	3
13791	1
13802	2
13807	3
13808	2
13809	1
13810	4
13822	4
13823	3
13825	4
13826	1
13833	3
13837	3
13842	4
13846	3
13852	5
13853	1
13858	1
13860	2
13866	2
13886	1
13887	0
13890	3
13891	3

INDEX	Predicted
13893	2
13902	6
13903	2
13908	4
13912	2
13924	1
13928	3
13929	4
13938	5
13939	2
13941	5
13951	2
13962	1
13964	2
13967	3
13971	3
13972	4
13975	0
13977	3
13979	4
13983	2
13984	3
13987	2
13994	4
13999	3
14003	6
14008	5
14011	5
14012	2
14016	3
14017	3
14020	1
14027	1
14038	4
14040	5
14042	3
14055	1
14057	5
14060	1
14081	1
14091	4
14111	3
14117	5
14121	4
14122	5
14125	2
14129	1
14135	1
14148	0
14157	4
14161	1
14163	1

INDEX	Predicted
14172	3
14180	1
14182	5
14188	1
14191	3
14201	7
14202	4
14213	4
14220	1
14224	2
14231	1
14241	5
14243	4
14245	3
14247	2
14248	3
14252	3
14254	4
14260	3
14269	5
14272	1
14274	4
14279	3
14280	5
14290	3
14298	7
14308	4
14313	4
14316	4
14319	6
14322	1
14323	3
14325	6
14337	4
14339	4
14341	3
14342	5
14346	6
14351	4
14354	1
14355	1
14358	6
14359	1
14364	4
14374	1
14376	2
14382	3
14384	3
14393	1
14398	3
14403	4
14406	0

INDEX	Predicted
14408	4
14411	1
14414	4
14418	3
14423	4
14442	4
14443	4
14444	3
14446	3
14455	3
14456	4
14458	4
14464	4
14466	5
14467	3
14469	3
14483	1
14484	3
14490	3
14491	2
14494	3
14496	4
14503	1
14504	3
14505	3
14506	1
14507	1
14512	1
14520	5
14527	2
14531	5
14532	1
14535	3
14543	1
14554	4
14556	7
14557	4
14561	2
14562	3
14567	1
14568	2
14574	3
14575	3
14579	3
14581	4
14582	2
14586	1
14591	3
14598	3
14599	3
14600	1
14612	5

INDEX	Predicted
14613	1
14624	5
14626	3
14630	4
14633	2
14639	3
14642	3
14643	4
14649	2
14650	2
14653	3
14655	2
14656	2
14662	4
14663	1
14673	3
14674	3
14676	4
14682	2
14685	5
14689	3
14693	3
14697	3
14700	3
14704	1
14710	3
14719	3
14724	5
14728	4
14735	3
14736	2
14741	1
14744	1
14753	1
14756	4
14762	5
14765	5
14783	5
14784	2
14786	0
14790	2
14793	3
14796	6
14801	1
14807	1
14812	4
14815	5
14831	4
14833	6
14836	6
14856	4
14859	3

INDEX	Predicted
14861	4
14863	0
14865	1
14880	1
14881	4
14883	1
14884	5
14894	4
14896	4
14899	3
14900	3
14901	2
14906	4
14907	1
14915	5
14919	3
14926	6
14927	2
14933	3
14937	3
14939	3
14940	2
14943	1
14953	4
14954	3
14969	4
14999	8
15008	4
15009	4
15018	5
15023	4
15025	1
15034	4
15036	4
15051	2
15052	3
15064	4
15070	4
15074	5
15077	5
15081	4
15086	6
15093	1
15094	0
15103	2
15104	1
15110	1
15112	2
15115	6
15131	1
15139	3
15141	3

INDEX	Predicted
15148	1
15154	6
15156	3
15161	2
15167	3
15178	4
15205	4
15207	1
15222	3
15223	4
15225	4
15228	5
15239	3
15241	1
15246	1
15247	2
15249	3
15255	3
15257	1
15267	1
15277	3
15280	5
15289	4
15297	1
15302	1
15304	1
15312	1
15321	0
15325	1
15326	5
15333	5
15337	1
15338	3
15340	6
15342	2
15344	3
15347	1
15349	5
15355	3
15359	1
15366	1
15367	0
15368	2
15369	5
15380	4
15381	3
15387	0
15388	3
15389	0
15392	1
15400	3
15405	4

INDEX	Predicted
15407	4
15408	7
15411	3
15413	4
15418	5
15419	5
15421	3
15425	3
15436	4
15438	4
15440	4
15443	4
15460	3
15464	1
15465	3
15473	1
15475	2
15483	1
15494	5
15495	7
15498	3
15499	3
15500	1
15501	1
15510	2
15512	2
15516	3
15518	5
15519	4
15524	3
15527	1
15529	2
15530	2
15538	1
15539	3
15541	1
15546	2
15547	3
15548	1
15552	2
15556	5
15567	3
15572	2
15573	5
15574	4
15577	1
15579	4
15581	3
15589	2
15596	2
15598	5
15599	4

INDEX	Predicted
15605	6
15606	4
15608	4
15616	5
15618	1
15621	1
15626	1
15638	3
15639	2
15642	2
15644	3
15646	5
15649	4
15656	2
15659	0
15680	1
15686	4
15693	1
15697	5
15699	5
15701	5
15705	2
15714	3
15722	4
15728	4
15734	3
15752	2
15756	4
15760	4
15762	4
15767	3
15768	1
15773	5
15774	4
15781	3
15782	3
15784	7
15791	2
15796	5
15798	5
15806	1
15814	1
15819	3
15825	4
15826	5
15831	4
15835	6
15836	2
15839	5
15845	3
15858	1
15859	4

INDEX	Predicted
15876	5
15878	4
15880	1
15886	4
15888	2
15891	4
15900	2
15902	4
15904	3
15908	2
15910	2
15917	3
15919	5
15924	3
15927	3
15937	2
15946	3
15949	3
15957	5
15961	4
15964	3
15965	3
15966	1
15978	1
15983	1
15987	2
15988	3
15998	2
16004	3
16008	3
16011	1
16023	3
16024	3
16025	3
16048	6
16050	3
16051	2
16057	1
16059	5
16060	4
16075	4
16094	6
16096	5
16116	2
16118	1
16121	2
16122	4
16124	6
16125	4
16126	3
16130	4

Part 1. Data Exploration

```
library(ggplot2)
library(psych)
library(knitr)
library(tidyr)
library(alr3)
library(effects)
library(pscl) # needed for zero inflation functions
```

Data Summary

```
# read raw training data set
url <- "https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/wine-training-data.csv"
hw5 <- read.csv(url, stringsAsFactors = FALSE)

# rename first column to something intelligible
colnames(hw5)[1] <- "INDEX"

d <- describe(hw5[,-c(1)])
d$mean <- round(d$mean,0)
d$sd <- round(d$sd,0)
d$min <- round(d$min,0)
d$max <- round(d$max,0)
d$range <- round(d$range,0)
d$skew <- round(d$skew,0)
d$kurtosis <- round(d$kurtosis,0)
d <- d[,-c(1,6,7)]
kable(d,digits=0)
```

Histograms

```
# remove index column
hw5.t <- hw5[,-1]

# now plot histograms
ggplot(gather(hw5.t, cols, value), aes(x = value)) +
       geom_histogram() + facet_wrap(~cols, scales = 'free')
```

Part 2. Data Preparation

Apply imputation to Training data set

```
library(psych)
library(knitr)
library(ggplot2)
library(tidyr)

hw5 <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/wine-training-data.

# rename first column to something intelligible - comes through garbled from Github for some reason
colnames(hw5)[1] <- "INDEX"

summary(hw5)
```

residual sugar 616 NA's chlorides 638 NA's freeSO2 647 NA's TotalSO2 682 NA's pH 395 NA's sulfates 1210 NA's alcohol 653 NA's stars 3359 NA's - would leave these blank

Step 1: Assume the stars left blank have no stars - they sold far fewer cases of wine. So impute as 0

```
hw5.1 <- hw5
hw5.1$STARS[which(is.na(hw5.1$STARS))] <- 0
```

Step 2: Shift all distributions having negative values to positive using absolute values

```
####################################### Variables that shouldn't have negative values ###############
# how many Alcohol values are < 0?: 118
summary(hw5.1$Alcohol)
hw5.1$Alcohol <- abs(hw5$Alcohol)
summary(hw5.1$Alcohol)

# how many Sulphates values are < 0? 2361
summary(hw5.1$Sulphates)
hw5.1$Sulphates <- abs(hw5$Sulphates)
summary(hw5.1$Sulphates)

# how many TotalSulfurDioxide values are < 0? 2504
summary(hw5.1$TotalSulfurDioxide)
hw5.1$TotalSulfurDioxide <- abs(hw5$TotalSulfurDioxide)
summary(hw5.1$TotalSulfurDioxide)

# how many FreeSulfurDioxide values are < 0? 3036
summary(hw5.1$FreeSulfurDioxide)
hw5.1$FreeSulfurDioxide <- abs(hw5$FreeSulfurDioxide)
summary(hw5.1$FreeSulfurDioxide)

# how many Chlorides values are < 0? 3197
summary(hw5.1$Chlorides)
hw5.1$Chlorides <- abs(hw5$Chlorides)
summary(hw5.1$Chlorides)

# how many ResidualSugar values are < 0? 3136
summary(hw5.1$ResidualSugar)
hw5.1$ResidualSugar <- abs(hw5$ResidualSugar)
summary(hw5.1$ResidualSugar)
```

```
# how many CitricAcid values are < 0? 2966
summary(hw5.1$CitricAcid)
hw5.1$CitricAcid <- abs(hw5$CitricAcid)
summary(hw5.1$CitricAcid)

# how many VolatileAcidity values are < 0? 2827
summary(hw5.1$VolatileAcidity)
hw5.1$VolatileAcidity <- abs(hw5$VolatileAcidity)
summary(hw5.1$VolatileAcidity)

# how many FixedAcidity values are < 0? 1621
summary(hw5.1$FixedAcidity)
hw5.1$FixedAcidity <- abs(hw5$FixedAcidity)
summary(hw5.1$FixedAcidity)



############### AFTER histograms
# remove rows containing NA's
hw5.t <- hw5.1[complete.cases(hw5.1), ]

# remove index column
hw5.t <- hw5.t[,-1]

ggplot(gather(hw5.t, cols, value), aes(x = value)) +
       geom_histogram() + facet_wrap(~cols, scales = 'free')

# plot new correlation matrix
round(cor(na.omit(hw5.1)), 2)
```

Step 3: Build model for pH

```
#Begin by taking out nonrelated predictors
ph <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide)
summary(ph)

#Remove fixed acidity
ph1 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity)
summary(ph1)

#Remove residual sugar
ph2 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Residu
summary(ph2)

#Remove STARS
ph3 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Residu
summary(ph3)

#Remove Alcohol
ph4 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Residu
summary(ph4)

#Remove sulphates
```

```
ph5 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Resid
summary(ph5)

#Remove free SO2
ph6 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Resid
summary(ph6)

#Remove density
ph7 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Resid
summary(ph7)

#Remove VolatileAcidity
ph8 <- lm(data=hw5.1, pH~.-INDEX - LabelAppeal - CitricAcid - TotalSulfurDioxide - FixedAcidity - Resid
summary(ph8)

#Remove chlorides
ph10 <- lm(data=hw5.1, pH~AcidIndex)
summary(ph10)

#Impute Function
impute <- function (a, a.impute){
  ifelse (is.na(a), a.impute,a)
}

#Impute pH
pred.ph <- round(predict(ph10, hw5.1), digits=2)

ph.Imp <- impute(hw5.1$pH, pred.ph)

hw5.1$pH <- as.numeric(ph.Imp)

summary(hw5.1$pH)
```

Step 4: Build Model for Alcohol

```
alc <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity)
summary(alc)

#Remove Density
alc1 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density)
summary(alc1)

#Remove pH
alc2 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH)
summary(alc2)

#Remove Sulphates
alc3 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates)
summary(alc3)

#Remove FixedAcidity
alc4 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc4)
```

```
#Remove Chlorides
alc5 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc5)

#Remove CitricAcid
alc6 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc6)

#Remove FreeSulfurDioxide
alc7 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc7)

#Remove ResidualSugar AND remove TARGET
alc8 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc8)



#Impute
pred.alc <- round(predict(alc8, hw5.1), digits=1)

alc.Imp <- impute(hw5.1$Alcohol, pred.alc)

hw5.1$Alcohol <- as.numeric(alc.Imp)

summary(hw5.1$Alcohol)

#imputed all but 37

# round(cor(na.omit(hw5.1)), 2)
```

Step 5: Build model for Residual Sugar

```
rs <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs)

#Remove Alcohol
rs1 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs1)

#Remove CitricAcid
rs2 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs2)

#Remove Density
rs3 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs3)

#Remove LabelAppeal
rs4 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs4)

#Remove AcidIndex
```

```
rs5 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs5)

#Remove FixedAcidity
rs6 <- lm(data=hw5.1, ResidualSugar~. - INDEX - TARGET - VolatileAcidity - Chlorides - pH - Sulphates -
summary(rs6)

# no significant predictors so use median
hw5.1$ResidualSugar[is.na(hw5.1$ResidualSugar)] <- median(hw5.1$ResidualSugar, na.rm = T)

summary(hw5.1$ResidualSugar)
```

Step 6: Build a model for chlorides

```
cl <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol)
summary(cl)

#remove STARS
cl2 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS)
summary(cl2)

#remove VolatileAcidity
cl3 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl3)

#remove CitricAcid
cl4 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl4)

#remove LabelAppeal
cl5 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl5)

#remove free SO2
cl6 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl6)

#remove pH
cl7 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl7)

#remove Total So2
cl8 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl8)

#remove TARGET
cl9 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl9)

#Impute
pred.cl <- round(predict(cl9, hw5.1), digits=3)

cl.Imp <- impute(hw5.1$Chlorides, pred.cl)
```

```r
hw5.1$Chlorides <- as.numeric(cl.Imp)

summary(hw5.1$Chlorides)

#Imputed all but 44
```

Step 7: Build a model for Sulphates

```r
SO <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide)
summary(SO)

#remove STARS
SO1 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS)
summary(SO1)

#remove ResidualSugar
SO2 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO2)

#remove VolatileAcidity
SO3 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO3)

#remove Total So2
SO4 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO4)

#remove pH
SO5 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO5)

#remove Density
SO6 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO6)

#remove CitricAcid AND TARGET
SO7 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO7)

#Impute
pred.SO <- round(predict(SO7, hw5.1), digits=2)

SO.Imp <- impute(hw5.1$Sulphates, pred.SO)

hw5.1$Sulphates <- as.numeric(SO.Imp)

# 44 not imputed

summary(hw5.1$Sulphates)
```

Step 8: Build a model for Free SO2

```
fso <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso)

#remove STARS
fso1 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso1)

#remove Chlorides
fso2 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso2)

#remove ResidualSugar
fso3 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso3)

#remove LabelAppeal
fso4 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso4)

#remove Alcohol
fso5 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso5)

#remove VolatileAcidity
fso6 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso6)

#remove Total so2
fso7 <- lm(data=hw5.1, FreeSulfurDioxide~. -INDEX - FixedAcidity - CitricAcid - Density - pH - Sulphates
summary(fso7)

# remove AcidIndex
fso8 <- lm(data=hw5.1, FreeSulfurDioxide ~ TARGET)
summary(fso8)

# use median since we can't use TARGET

hw5.1$FreeSulfurDioxide[is.na(hw5.1$FreeSulfurDioxide)] <- median(hw5.1$FreeSulfurDioxide, na.rm = T)
```

Step 9: Build a model for Total SO2

```
tso <- lm(data=hw5.1, TotalSulfurDioxide~.-INDEX)
summary(tso)

#Remove FixedAcidity
tso1 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity)
summary(tso1)

#Remove Sulphates
tso2 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates)
summary(tso2)

#Remove STARS
```

```r
tso3 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS)
summary(tso3)

#Remove CitricAcid
tso4 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid)
summary(tso4)

#Remove pH
tso5 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso5)

#Remove ResidualSugar
tso6 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso6)

#Remove FreeSulfurDioxide
tso7 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso7)

#Remove Chlorides
tso8 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso8)

#Remove Density AND Target
tso9 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso9)

# remove LabelAppeal
tso10 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - 
summary(tso10)


#Impute
pred.tso <- round(predict(tso10, hw5.1), digits=0)

tso.Imp <- impute(hw5.1$TotalSulfurDioxide, pred.tso)

hw5.1$TotalSulfurDioxide <- as.numeric(tso.Imp)

summary(hw5.1$TotalSulfurDioxide)

#imputed all but 37
```

Step 10: Impute remaining missing values

```r
#Chlorides - Impute median as it needs the free SO2 which is missing
hw5.1$Chlorides[is.na(hw5.1$Chlorides)] <- median(hw5.1$Chlorides, na.rm = T)
#Chlorides is complete

#Alcohol - IMpute median as it needs the free SO2 which is missing
hw5.1$Alcohol[is.na(hw5.1$Alcohol)] <- median(hw5.1$Alcohol, na.rm = T)
#Alcohol is complete
```

```r
#NOw that alcohol and chlorides are complete, can repeat model for Total SO2 and Sulphates
pred.tso <- round(predict(tso10, hw5.1), digits=0)
tso.Imp <- impute(hw5.1$TotalSulfurDioxide, pred.tso)
hw5.1$TotalSulfurDioxide <- as.numeric(tso.Imp)

pred.sulph <- round(predict(S07, hw5.1), digits=0)
sulph.Imp <- impute(hw5.1$Sulphates, pred.sulph)
hw5.1$Sulphates <- as.numeric(sulph.Imp)

#Compare before and after

describe(hw5)
describe(hw5.1)

# summary(hw5)
# summary(hw5.1)

############### AFTER histograms
# remove rows containing NA's
hw5.t <- hw5.1[complete.cases(hw5.1), ]

# remove index column
hw5.t <- hw5.t[,-1]

ggplot(gather(hw5.t, cols, value), aes(x = value)) +
       geom_histogram() + facet_wrap(~cols, scales = 'free')
```

Write output to a CSV file

```r
write.csv(hw5.1, file = "C:/SQLData/621/HW5-IMPUTED-DATA.csv", row.names = FALSE)

# now do a read test

hw5.test <- read.csv("C:/SQLData/621/HW5-IMPUTED-DATA.csv", stringsAsFactors = FALSE)

summary(hw5.test)
```

Apply imputation to evaluation data set

```r
library(psych)
library(knitr)
library(ggplot2)
library(tidyr)

hw5 <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/wine-evaluation-data

# rename first column to something intelligible - comes through garbled from Github for some reason
colnames(hw5)[1] <- "INDEX"

summary(hw5)
```

residual sugar 616 NA's chlorides 638 NA's freeSO2 647 NA's TotalSO2 682 NA's pH 395 NA's sulfates 1210 NA's alcohol 653 NA's stars 3359 NA's - would leave these blank

Step 1: Assume the stars left blank have no stars - they sold far fewer cases of wine. So impute as 0

```
hw5.1 <- hw5
hw5.1$STARS[which(is.na(hw5.1$STARS))] <- 0

# Set TARGET = 0

hw5.1$TARGET <- 0
```

Step 2: Shift all distributions having negative values to positive using + abs(min(x))

```
######################################## Variables that shouldn't have negative values ###############
# how many Alcohol values are < 0?: 118
summary(hw5.1$Alcohol)
hw5.1$Alcohol <- abs(hw5$Alcohol)
summary(hw5.1$Alcohol)

# how many Sulphates values are < 0? 2361
summary(hw5.1$Sulphates)
hw5.1$Sulphates <- abs(hw5$Sulphates)
summary(hw5.1$Sulphates)

# how many TotalSulfurDioxide values are < 0? 2504
summary(hw5.1$TotalSulfurDioxide)
hw5.1$TotalSulfurDioxide <- abs(hw5$TotalSulfurDioxide)
summary(hw5.1$TotalSulfurDioxide)

# how many FreeSulfurDioxide values are < 0? 3036
summary(hw5.1$FreeSulfurDioxide)
hw5.1$FreeSulfurDioxide <- abs(hw5$FreeSulfurDioxide)
summary(hw5.1$FreeSulfurDioxide)

# how many Chlorides values are < 0? 3197
summary(hw5.1$Chlorides)
hw5.1$Chlorides <- abs(hw5$Chlorides)
summary(hw5.1$Chlorides)

# how many ResidualSugar values are < 0? 3136
summary(hw5.1$ResidualSugar)
hw5.1$ResidualSugar <- abs(hw5$ResidualSugar)
summary(hw5.1$ResidualSugar)


# how many CitricAcid values are < 0? 2966
summary(hw5.1$CitricAcid)
hw5.1$CitricAcid <- abs(hw5$CitricAcid)
summary(hw5.1$CitricAcid)

# how many VolatileAcidity values are < 0? 2827
summary(hw5.1$VolatileAcidity)
hw5.1$VolatileAcidity <- abs(hw5$VolatileAcidity)
```

```r
summary(hw5.1$VolatileAcidity)

# how many FixedAcidity values are < 0? 1621
summary(hw5.1$FixedAcidity)
hw5.1$FixedAcidity <- abs(hw5$FixedAcidity)
summary(hw5.1$FixedAcidity)



############### AFTER histograms
# remove rows containing NA's
hw5.t <- hw5.1[complete.cases(hw5.1), ]

# remove index column
hw5.t <- hw5.t[,-1]

ggplot(gather(hw5.t, cols, value), aes(x = value)) +
       geom_histogram() + facet_wrap(~cols, scales = 'free')

# plot new correlation matrix
round(cor(na.omit(hw5.1)), 2)
```

Step 3: Build model for pH

```r
ph10 <- lm(data=hw5.1, pH~AcidIndex)
summary(ph10)

#Impute Function
impute <- function (a, a.impute){
  ifelse (is.na(a), a.impute,a)
}

#Impute pH
pred.ph <- round(predict(ph10, hw5.1), digits=2)

ph.Imp <- impute(hw5.1$pH, pred.ph)

hw5.1$pH <- as.numeric(ph.Imp)

summary(hw5.1$pH)
```

Step 4: Build Model for Alcohol

```r
#Remove ResidualSugar AND remove TARGET
alc8 <- lm(data=hw5.1, Alcohol~. - INDEX - LabelAppeal - VolatileAcidity - Density - pH - Sulphates - F
summary(alc8)

#Impute
pred.alc <- round(predict(alc8, hw5.1), digits=1)

alc.Imp <- impute(hw5.1$Alcohol, pred.alc)

hw5.1$Alcohol <- as.numeric(alc.Imp)
```

```
summary(hw5.1$Alcohol)

# Some NA's

# round(cor(na.omit(hw5.1)), 2)
```

Step 5: Build model for Residual Sugar

```
# no significant predictors so use median
hw5.1$ResidualSugar[is.na(hw5.1$ResidualSugar)] <- median(hw5.1$ResidualSugar, na.rm = T)

summary(hw5.1$ResidualSugar)

# round(cor(na.omit(hw5.1)), 2)
```

Step 6: Build a model for chlorides

```
#remove TARGET
cl9 <- lm(data=hw5.1, Chlorides~.-INDEX - FixedAcidity - ResidualSugar - Alcohol - STARS - VolatileAcidi
summary(cl9)

#Impute
pred.cl <- round(predict(cl9, hw5.1), digits=3)

cl.Imp <- impute(hw5.1$Chlorides, pred.cl)

hw5.1$Chlorides <- as.numeric(cl.Imp)

summary(hw5.1$Chlorides)

# Some NA's

# round(cor(na.omit(hw5.1)), 2)
```

Step 7: Build a model for Sulphates

```
#remove CitricAcid AND TARGET
SO7 <- lm(data=hw5.1, Sulphates~.-INDEX - LabelAppeal -Alcohol - FreeSulfurDioxide - STARS - ResidualSug
summary(SO7)

#Impute
pred.SO <- round(predict(SO7, hw5.1), digits=2)

SO.Imp <- impute(hw5.1$Sulphates, pred.SO)

hw5.1$Sulphates <- as.numeric(SO.Imp)

# Some NA's

summary(hw5.1$Sulphates)

# round(cor(na.omit(hw5.1)), 2)
```

Step 8: Build a model for Free SO2

```
# use median since we can't use TARGET

hw5.1$FreeSulfurDioxide[is.na(hw5.1$FreeSulfurDioxide)] <- median(hw5.1$FreeSulfurDioxide, na.rm = T)
```

Step 9: Build a model for Total SO2

```
# remove LabelAppeal
tso10 <- lm(data=hw5.1, TotalSulfurDioxide~. - INDEX - FixedAcidity - Sulphates - STARS - CitricAcid - p
summary(tso10)


#Impute
pred.tso <- round(predict(tso10, hw5.1), digits=0)

tso.Imp <- impute(hw5.1$TotalSulfurDioxide, pred.tso)

hw5.1$TotalSulfurDioxide <- as.numeric(tso.Imp)

summary(hw5.1$TotalSulfurDioxide)

# Some NA's
```

Step 10: Impute remaining missing values

```
#Chlorides - Impute median as it needs the free SO2 which is missing
hw5.1$Chlorides[is.na(hw5.1$Chlorides)] <- median(hw5.1$Chlorides, na.rm = T)
#Chlorides is complete

#Alcohol - IMpute median as it needs the free SO2 which is missing
hw5.1$Alcohol[is.na(hw5.1$Alcohol)] <- median(hw5.1$Alcohol, na.rm = T)
#Alcohol is complete

#NOw that alcohol and chlorides are complete, can repeat model for Total SO2 and Sulphates
pred.tso <- round(predict(tso10, hw5.1), digits=0)
tso.Imp <- impute(hw5.1$TotalSulfurDioxide, pred.tso)
hw5.1$TotalSulfurDioxide <- as.numeric(tso.Imp)

pred.sulph <- round(predict(SO7, hw5.1), digits=0)
sulph.Imp <- impute(hw5.1$Sulphates, pred.sulph)
hw5.1$Sulphates <- as.numeric(sulph.Imp)

#Compare before and after

describe(hw5)
describe(hw5.1)

# summary(hw5)
# summary(hw5.1)

############### AFTER histograms
# remove rows containing NA's
```

```
hw5.t <- hw5.1[complete.cases(hw5.1), ]

# remove index column
hw5.t <- hw5.t[,-1]

ggplot(gather(hw5.t, cols, value), aes(x = value)) +
       geom_histogram() + facet_wrap(~cols, scales = 'free')
```

Write output to a CSV file

```
write.csv(hw5.1, file = "C:/SQLData/621/HW5-IMPUTED-EVAL-DATA.csv", row.names = FALSE)

# now do a read test

hw5.test <- read.csv("C:/SQLData/621/HW5-IMPUTED-EVAL-DATA.csv", stringsAsFactors = FALSE)

summary(hw5.test)
```

Part 3. Build Models

```
# read imputed data set
hw5 <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/HW5-IMPUTED-DATA.cs
```

Poisson Model 1

```
pois.1 <- glm(TARGET ~ . - INDEX, family=poisson, hw5)
summary(pois.1)

# plot residuals vs fitted
plot(round(predict(pois.1)),residuals(pois.1),xlab="Fitted",ylab="
Residuals")


# check plot to see if an outlier is causing the poor fit
halfnorm(residuals(pois.1))

# plot estimated variance against the mean,
#plot(log(fitted(pois.1)),log((hw5$TARGET-fitted(pois.1))^2),xlab=expression(hat(mu)),ylab=expression((
#abline(0,1)

# estimate the dispersion parameter: 0.8529 so since very close to 1 its OK
(dp <- sum(residuals(pois.1,type="pearson")^2)/pois.1$df.res)

# check summary to see if dropping a predictor will improve things
```

```r
# the p-values get adjusted by the dispersion so re-check them!!!
summary(pois.1,dispersion=dp)

# another check to see if dropping a predictor will improve things: check for low F-scores
drop1(pois.1, test="F")

# -----------------------------------
# Both sets of results say drop FixedAcidity, ResidualSugar. CitricAcid

# REFIT
pois.2 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid, family=poisson, hw5)
summary(pois.2)


# remove FreeSulfurDioxide
pois.3 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide, family=
summary(pois.3)

# remove Density
pois.4 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide - Densit
summary(pois.4)

# remove Chlorides
pois.5 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide - Densit
summary(pois.5)

# remove Alcohol
pois.6 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide - Densit
summary(pois.6)


# check diagnostic plots
par(mfrow = c(2,2))
plot(pois.6)

# check mmp's
mmps(pois.6,layout=c(3,3),key=TRUE)

# check plot to see if an outlier is causing the poor fit
halfnorm(residuals(pois.2))

# estimate the dispersion parameter: is same as above: 0.8512 so its OK
(dp <- sum(residuals(pois.6,type="pearson")^2)/pois.6$df.res)

# check summary to see if dropping a predictor will improve things
summary(pois.6,dispersion=dp)
drop1(pois.6, test="F")
# results say drop sulfates and pH

# remove Sulphates + pH
pois.7 <- glm(TARGET ~ . - INDEX- FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide - Densit
summary(pois.7)
```

```
# proportion of deviance explained by the model:

# 1 - (14774/22861)

drop1(pois.7, test="F")

# STOP

# check diagnostic plots
par(mfrow = c(2,2))
plot(pois.7)

# check mmp's
# mmps(pois.7,layout=c(3,3),key=TRUE)

# check plot to see if an outlier is causing the poor fit
halfnorm(residuals(pois.7))

# estimate the dispersion parameter: is same as above: 0.8537 so its OK
(dp <- sum(residuals(pois.7,type="pearson")^2)/pois.7$df.res)

# check goodness of fit
1 - pchisq(deviance(pois.7),df.residual(pois.7))

# anova goodness of fit: indicates good fit for all variables
anova(pois.7, test="Chisq")

# influential / leverage points - basically an outlier plot
influencePlot(pois.7)
```

Outliers found at 723, 8887, 11289 Remove and refit

```
############ FIRST SET OF OUTLIERS ######################
# drop outlier records from data set
hw5p_rem <- hw5[-c(723, 8887, 11289), ]

# renumber rows
rownames(hw5p_rem) <- 1:nrow(hw5p_rem)

pois.1 <- glm(TARGET ~ . - INDEX, family=poisson, hw5p_rem)
summary(pois.1)
drop1(pois.1, test="F")

# drop says remove FixedAcidity, CitricAcid, ResidualSugar, Chlorides, FreeSulfurDioxide, Density,
# pH, Sulphates, Alcohol

pois.2 <- glm(TARGET ~ . - INDEX - FixedAcidity - ResidualSugar - CitricAcid - Density - Chlorides - Fr
summary(pois.2)
drop1(pois.2, test="F")

# check diagnostic plots
par(mfrow = c(2,2))
plot(pois.2)
```

```r
mmps(pois.2, labels = TRUE)

# influential / leverage points - basically an outlier plot
influencePlot(pois.2)

# check plot to see if an outlier is causing the poor fit
halfnorm(residuals(pois.2))

# estimate the dispersion parameter: is same as above: 0.8524 so its OK
(dp <- sum(residuals(pois.2,type="pearson")^2)/pois.2$df.res)

# check goodness of fit
1 - pchisq(deviance(pois.2),df.residual(pois.2))

# anova goodness of fit: indicates good fit to all variables
anova(pois.2, test="Chisq")

plot(allEffects(pois.2, default.levels=50), ask = F)

# now plot TARGET against fitted values
fit1 <- round(pois.2$fitted.values)
par(mfrow = c(1,1))

plot(fit1, hw5p_rem$TARGET, xlab="Fitted Values")
abline(lsfit(fit1, hw5p_rem$TARGET),lty=2)


res <- residuals(pois.2, type="response")
plot(log(predict(pois.2)), res)
abline(h=0, lty=2)
qqnorm(res)
qqline(res)
```

Poisson Model 2: Zero Inflated Poisson

```r
pois.zinf <- zeroinfl(TARGET ~  VolatileAcidity + TotalSulfurDioxide + LabelAppeal + AcidIndex +
                       STARS, data= hw5p_rem, dist="poisson")
# check zero inflated model coefficients
summary(pois.zinf)

# test models based on AIC scores - zero inflated is better
AIC(pois.2, pois.zinf)

# test models based on Vuong test - zero inflate is better
vuong(pois.2, pois.zinf)

# check dispersion: it has decreased from 0.85 to 0.46
(dp <- sum(residuals(pois.zinf,type="pearson")^2)/pois.zinf$df.res)
```

```
# compare to null model
mnull <- update(pois.zinf, . ~ 1)

# 5 predictors so we have 8 df (n-2) for pchisq test
pchisq(2 * (logLik(pois.zinf) - logLik(mnull)), df = 3, lower.tail = FALSE)
# result of p = 0 shows zero inflated model is much more accurate

res <- residuals(pois.zinf, type="response")
plot(log(predict(pois.zinf)), res)
abline(h=0, lty=2)
qqnorm(res)
qqline(res)

# now plot TARGET against fitted values
fit1 <- round(pois.zinf$fitted.values)
par(mfrow = c(1,1))

plot(fit1, hw5$TARGET, xlab="Fitted Values")
abline(lsfit(fit1, hw5$TARGET),lty=2)
```

Negative Binomial Model 1

```
library(MASS)

# modn <- glm(skips ~ .,negative.binomial(1),solder)
# modn

# Better way to run negative binomial is to allow the parameter k to vary and be
# estimated using maximum likelihood. USE THIS FUNCTION INSTEAD
negb.1 <- glm.nb(TARGET ~ . - INDEX, hw5)
summary(negb.1)

# calculate dispersion: is identical to Poisson = 0.8512
(dp <- sum(residuals(negb.1,type="pearson")^2)/negb.1$df.res)

# check summary to see if dropping a predictor will improve things
# the p-values get adjusted by the dispersion so re-check them!!!
summary(negb.1,dispersion=dp)

# another check to see if dropping a predictor will improve things: check for low F-scores
drop1(negb.1, test="F")

# Both sets of results say drop FixedAcidity, ResidualSugar CitricAcid, Chlorides, FreeSulfurDioxide
# Density, Sulphates, Alcohol, pH
# so refit
negb.2 <- glm.nb(TARGET ~ . - INDEX - FixedAcidity - ResidualSugar - CitricAcid - Density - Chlorides -
summary(negb.2)

influencePlot(negb.2)
```

```
############ FIRST SET OF OUTLIERS ######################
# drop outlier records from data set
hw5p_rem <- hw5[-c(723, 8887, 11289), ]

# renumber rows
rownames(hw5p_rem) <- 1:nrow(hw5p_rem)

# refit model
negb.3 <- glm.nb(TARGET ~ . - INDEX - FixedAcidity - ResidualSugar - CitricAcid - Density - Chlorides -
summary(negb.3)

(dp <- sum(residuals(negb.3,type="pearson")^2)/negb.3$df.res)

influencePlot(negb.3)

# check goodness of fit
1 - pchisq(deviance(negb.3),df.residual(negb.3))

# anova goodness of fit: indicates good fit to all variables
anova(negb.3, test="Chisq")

# check diagnostic plots
par(mfrow = c(2,2))
plot(negb.3)

plot(allEffects(negb.3, default.levels=50), ask = F)

# now plot TARGET against fitted values
fit1 <- round(negb.3$fitted.values)
par(mfrow = c(1,1))

plot(fit1, hw5$TARGET, xlab="Fitted Values")
abline(lsfit(fit1, hw5$TARGET),lty=2)
```

Negative Binomial Model 2: Zero inflated

```
negb.zinf <- zeroinfl(TARGET ~ VolatileAcidity + TotalSulfurDioxide + LabelAppeal + AcidIndex +
                       STARS, data=
                        hw5p_rem, dist="negbin", EM = TRUE)

# check zero inflated model coefficients
summary(negb.zinf)

AIC(negb.3, negb.zinf)

# compare to null model
mnull <- update(negb.zinf, . ~ 1)

# 5 predictors so we have 3 df (n-2) for pchisq test
```

```
pchisq(2 * (logLik(negb.zinf) - logLik(mnull)), df = 3, lower.tail = FALSE)
# result of p = 0 shows zero inflated model is much more accurate

# now plot TARGET against fitted values
fit1 <- round(negb.zinf$fitted.values)
par(mfrow = c(1,1))

plot(fit1, hw5p_rem$TARGET, xlab="Fitted Values")
abline(lsfit(fit1, hw5p_rem$TARGET),lty=2)
```

Linear Model 1

Load Imputed Dataset

```
url <-"https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/HW5-IMPUTED-DATA.csv"
hw5.t <- read.csv(url, stringsAsFactors = FALSE)
```

Add STARS

```
modl.1 <- lm(TARGET ~ STARS, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.1)

#DIAGNOSTIC1. can't show collinearity w/ 1st variable. Check p-value to < 0.05.
#p-values are all < 0.05

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.1, ~., ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.1)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.1)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
```

```
abline(h=0)
plot(modl.1$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.1$fitted.values, modl.1$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.1$fitted.values,hw5.t$TARGET))
plot(modl.1)
# normal distribution, and uniform distribution of residuals
```

Add LabelAppeal

```
modl.2 <- lm(TARGET ~ STARS + LabelAppeal, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.2)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.2)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.2, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.2)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.2)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.2$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals
```

```
#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.2$fitted.values, modl.2$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.2$fitted.values,hw5.t$TARGET))
plot(modl.2)
# normal distribution, and uniform distribution of residuals
```

Add AcidIndex

```
modl.3 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.3)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.3)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.3, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.3)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.3)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.3$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.3$fitted.values, modl.3$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.3$fitted.values,hw5.t$TARGET))
```

```
plot(modl.3)
# normal distribution, and uniform distribution of residuals
```

Add VolatileAcidity

```
modl.4 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.4)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.4)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.4, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.4)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.4)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$VolatileAcidity,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.4$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.4$fitted.values, modl.4$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.4$fitted.values,hw5.t$TARGET))
```

```
plot(modl.4)
# normal distribution, and uniform distribution of residuals
```

Add Alcohol

```
modl.5 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.5)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.5)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.5, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.5)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.5)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$VolatileAcidity,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Alcohol,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.5$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
```

```
plot(modl.5$fitted.values, modl.5$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.5$fitted.values,hw5.t$TARGET))
plot(modl.5)
# normal distribution, and uniform distribution of residuals
```

Add FixedAcidity

```
modl.6 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + FixedAcidity, hw5.t

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.6)

#p-values are all < 0.05 and no VIFs  > 5
```

Remove FixedAcidity due to high p-value Add Density

```
modl.7 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + Density, hw5.t)

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.7)
```

Remove Density due to high p-value Add TotalSulfurDioxide

```
modl.8 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide,

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.8)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.8)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.8, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.8)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
```

```
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.5)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$VolatileAcidity,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Alcohol,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$TotalSulfurDioxide,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.5$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.8$fitted.values, modl.8$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.8$fitted.values,hw5.t$TARGET))
plot(modl.8)
```

Add Chlorides

```
modl.9 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide +

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.9)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.9)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.9, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.9)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
```

```
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.5)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$VolatileAcidity,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Alcohol,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$TotalSulfurDioxide,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Chlorides,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.5$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.9$fitted.values, modl.9$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.9$fitted.values,hw5.t$TARGET))
plot(modl.9)
```

Add FreeSulfurDioxide

```
modl.10 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.10)
```

Remove Sulphates due to high p-value Add FreeSulfurDioxide

```
modl.11 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.11)
```

Remove FreeSulfurDioxide due to high p-value Add CitricAcid

```
modl.12 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.12)
```

Remove CitricAcid due to high p-value Add pH

```
modl.13 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.13)

#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(modl.13)
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(modl.13, ~.,ask=FALSE, id.n = 2)
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(modl.13)
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

##DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(modl.5)
plot(hw5.t$STARS,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$LabelAppeal,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$AcidIndex,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$VolatileAcidity,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Alcohol,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$TotalSulfurDioxide,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(hw5.t$Chlorides,StanRest,ylab="Standardized Residuals")
```

```
abline(h=0)
plot(hw5.t$pH,StanRest,ylab="Standardized Residuals")
abline(h=0)
plot(modl.5$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
abline(h=0)
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(modl.13$fitted.values, modl.13$TARGET,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(modl.13$fitted.values,hw5.t$TARGET))
plot(modl.13)
```

Don't add ResidualSugar due to high p-value

```
modl.14 <- lm(TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity + Alcohol + TotalSulfurDioxide

# generate standard residuals
par(mfrow=c(1,3))

summary(modl.14)
```

Linear Model 2

```
hw5i <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/HW5-IMPUTED-DATA.cs
par(mfrow=c(1,1))
library(MASS)
library(effects)
library(car)


lmod <- lm(data=hw5i, TARGET~.- INDEX)
summary(lmod, dispersion =dp)
library(effects)
drop1(lmod, test="F")

#remove fixed acidity
lmod1 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity)
summary(lmod1)
drop1(lmod1, test="F")

#remove residual sugar
lmod2 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar)
summary(lmod2)
drop1(lmod2, test="F")

#remove free SO2
lmod3 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar - FreeSulfurDioxide)
```

```r
summary(lmod3)
drop1(lmod3, test="F")

#remove citric acid
lmod4 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide)
summary(lmod4)
drop1(lmod4, test="F")

#remove chlorides
lmod5 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide
summary(lmod5)
drop1(lmod5, test="F")

#remove density
lmod6 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide
summary(lmod6)
drop1(lmod6, test="F")

#remove pH
lmod7 <- lm(data=hw5i, TARGET~.- INDEX - FixedAcidity - ResidualSugar - CitricAcid - FreeSulfurDioxide
summary(lmod7)
plot(allEffects(lmod7, default.levels=50), ask = F)
#residual
library(faraway)
halfnorm(residuals(lmod7))
#shows no outliers
dp7 <- sum(residuals(lmod7, type="pearson")^2)/lmod7$df.residual
dp7
#1.7605


#look at power transformations
library(car)
library(MASS)
summary(powerTransform(LabelAppeal+3~TARGET, hw5i, family="bcPower"))
boxcox(hw5i$LabelAppeal +3~hw5i$TARGET)
#no transformation

summary(powerTransform(STARS +1~TARGET, hw5i, family="bcPower"))
boxcox(hw5i$STARS +1~hw5i$TARGET)
#sqrt

summary(powerTransform(Alcohol + 1~TARGET, hw5i, family="bcPower"))
boxcox(hw5i$Alcohol +1~hw5i$TARGET)
#no transformation

#transformed variables
lmod8 <- lm(data=hw5i, TARGET~VolatileAcidity + TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal
summary(lmod8)
dp8 <- sum(residuals(lmod8, type="pearson")^2)/lmod8$df.residual
dp8
#1.735
```

```
#remove sulphates
lmod9 <- lm(data=hw5i, TARGET~VolatileAcidity + TotalSulfurDioxide + Alcohol + LabelAppeal + AcidIndex +
summary(lmod9)
dp.9 <- sum(residuals(lmod9, type="pearson")^2)/lmod9$df.residual
dp.9

#STARS definitely improved  ADJ R2 = 0.5323
plot(lmod9)
mmps(lmod9,layout=c(3,4),key=TRUE)
```

Part 4. Select Models

R code for the required 2-stage prediction process

First stage:

1) Load post-imputation training data set
2) build selected regression model on full training data set
3) randomly select 100 rows
4) Apply regression model using **predict** function to 100 rows and compare results to actual TARGET value
5) Save predictions and 100 row subset to a file.

```
library(knitr)
# library(faraway) # needed for some diagnostic functions
library(pscl) # needed for zero inflation functions
library(alr3)
library(car)
library(psych)
options(scipen=999)

# load training set so that binary model can be built

# read imputed data
hw5 <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/HW5-IMPUTED-DATA.csv

############ FIRST SET OF OUTLIERS ######################
# drop outlier records from data set
hw5p_rem <- hw5[-c(723, 8887, 11289), ]

# renumber rows
rownames(hw5p_rem) <- 1:nrow(hw5p_rem)

# save a copy of TARGET_AMT for stats at end
Target.amt <- hw5$TARGET

# build model
negb.zinf <- zeroinfl(TARGET ~ VolatileAcidity + TotalSulfurDioxide + LabelAppeal + AcidIndex +
                       STARS, data=
                        hw5p_rem, dist="negbin", EM = TRUE)
```

```
# sample 100 rows
set.seed(123)
mtest <- hw5[sample(nrow(hw5), 100), ]

# now predict TARGET_FLAG using model
pred.w <- round(predict(negb.zinf, newdata=mtest, type="response"))
summary(pred.w)

# create a dataframe containing only the relevant items
mtest.out <- cbind(mtest$INDEX, mtest$TARGET, data.frame(pred.w))

# rename columns
colnames(mtest.out) <- c("INDEX","TARGET","Predicted")

# compare metrics
describe(mtest.out$Predicted)
describe(mtest.out$TARGET)

# plot response vs fitted
plot(jitter(pred.w),  mtest$TARGET, xaxp  = c(-1, 10, 11))

# predicted values have lower variance but similar mean

# write results to a file

write.csv(mtest.out, file = "C:/SQLData/621/HW5-PRED-TEST-DATA.csv", row.names = FALSE)
```

Second stage:

- Apply selected model to evaluation data set, generate comparative statistics, and save predictions to a file

```
# ----------------------------------------

# now that model is built, load eval data set

# load EVAL data set
hw5 <- read.csv("https://raw.githubusercontent.com/jtopor/CUNY-MSDA-621/master/HW-5/HW5-IMPUTED-EVAL-DA

# now predict TARGET_FLAG using model
pred.w <- round(predict(negb.zinf, newdata=hw5, type="response"))
summary(pred.w)

# create a dataframe containing only the relevant items
eval.out <- cbind(hw5$INDEX, data.frame(pred.w))

# rename columns
colnames(eval.out) <- c("INDEX", "Predicted")

# compare metrics
describe(eval.out$Predicted)
describe(Target.amt)
```

```
write.csv(eval.out, file = "C:/SQLData/621/HW5-PRED-EVAL-COLS-ONLY.csv", row.names = FALSE)
```