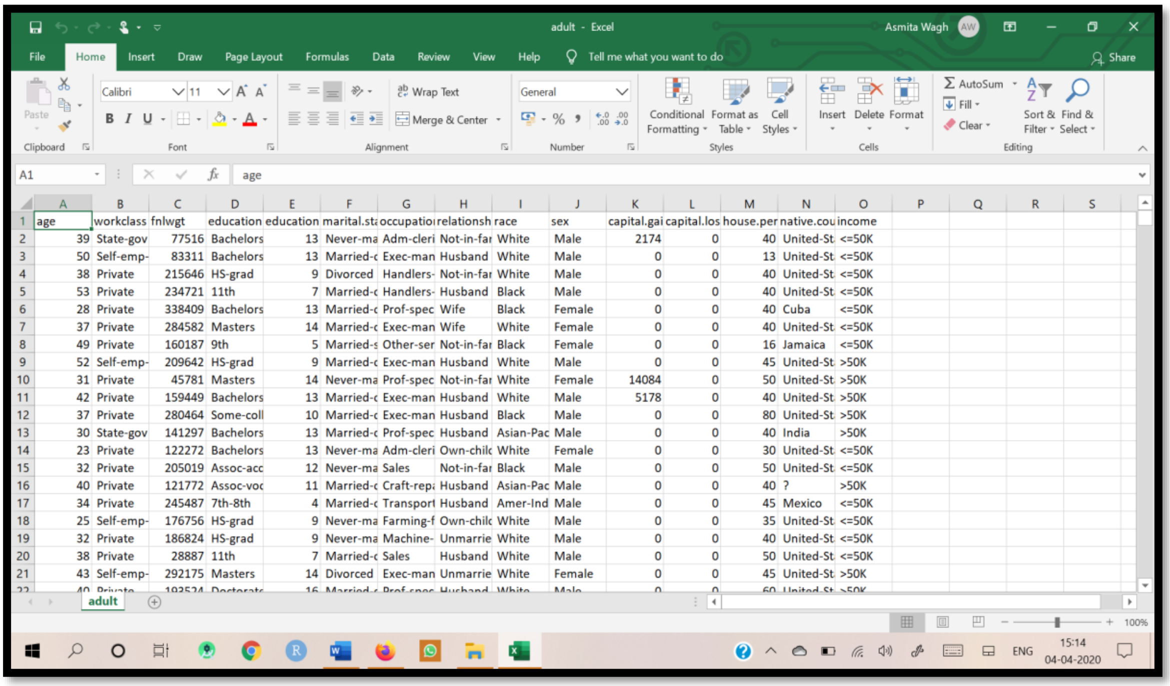
Experiment no:12

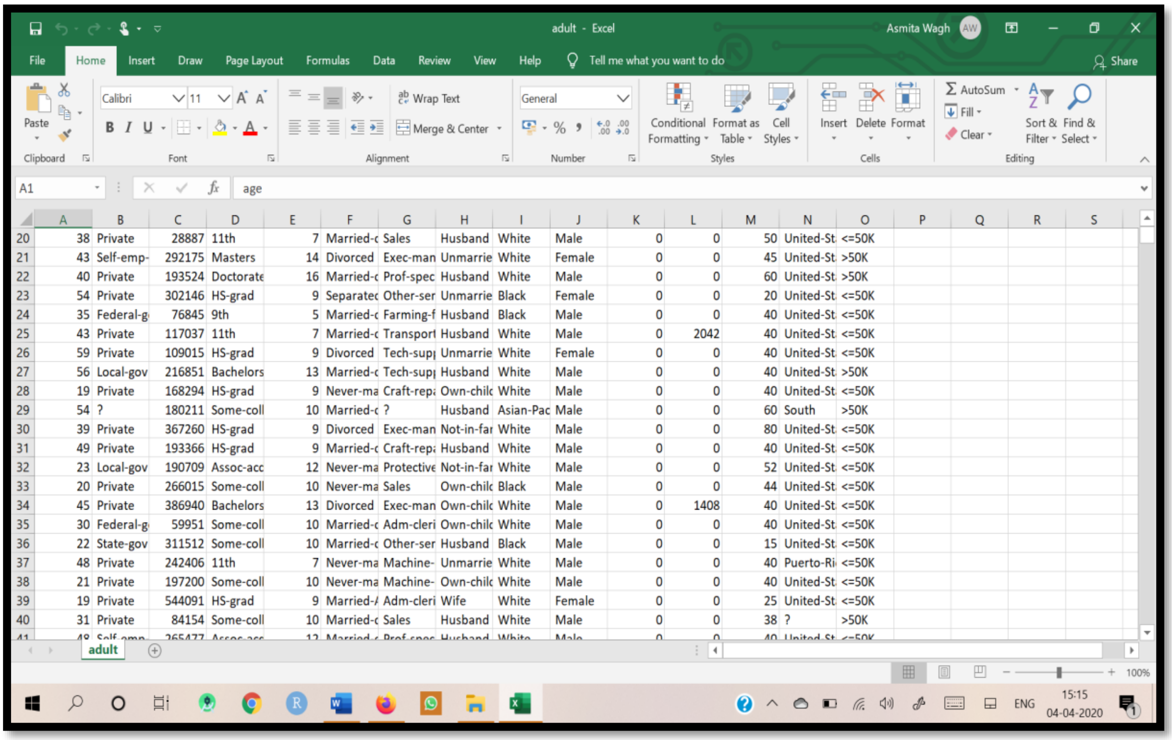
Project: Mini-Project for R-Lab

Project-Name: Salary Prediction

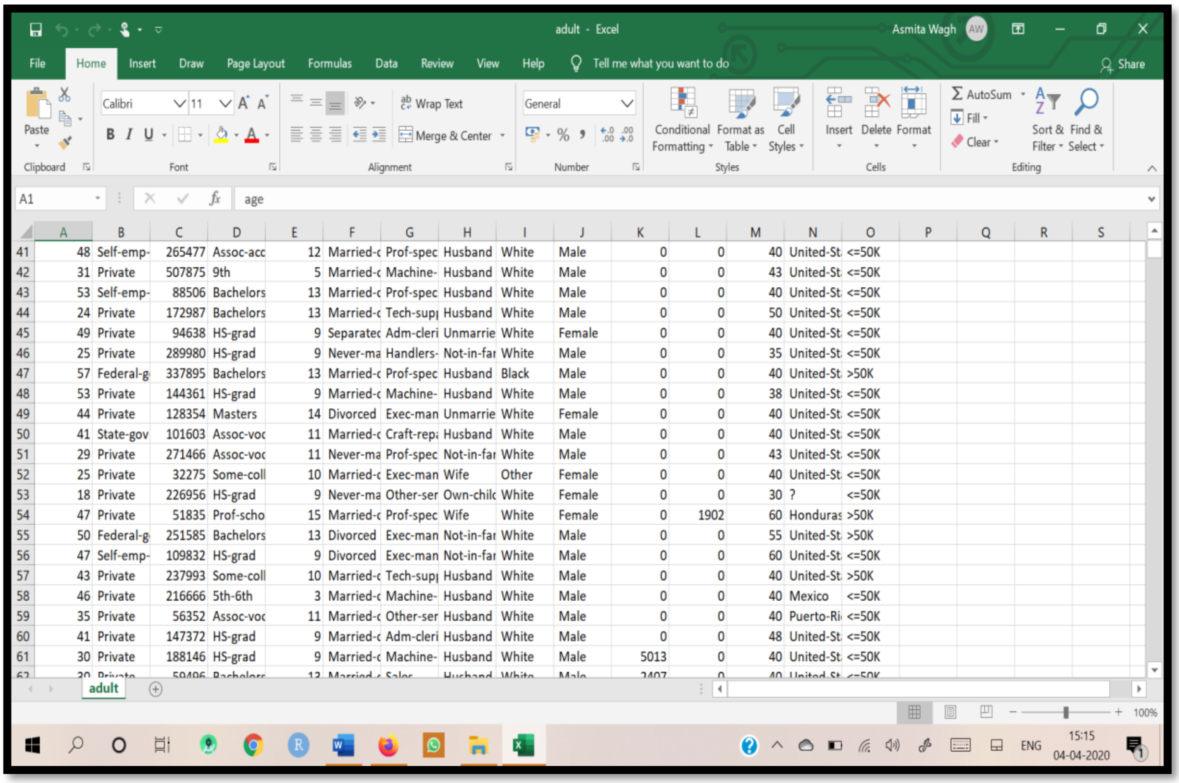
Dataset:



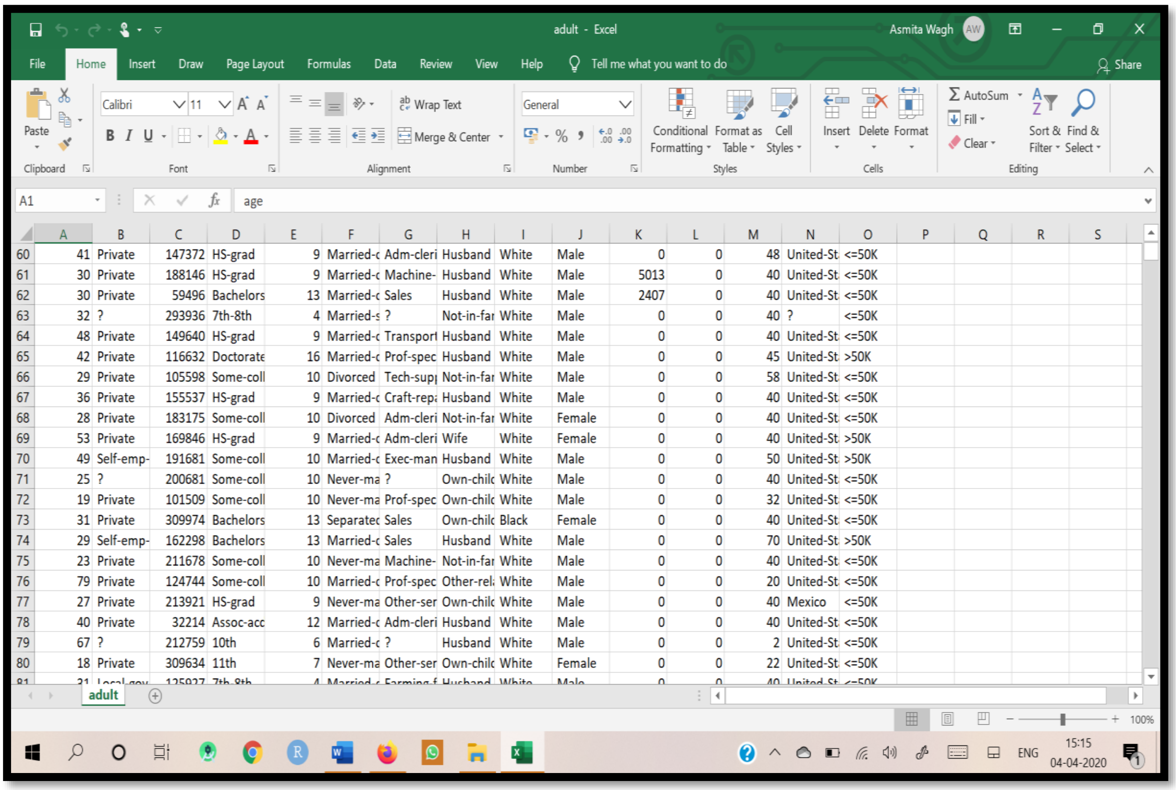
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	workclass	fnlwgt	education	education	marital.sta	occupatio	relationsh	race	sex	capital.gai	capital.los	house.per	native.cou	income
2	39	State-gov	77516	Bachelors	13	Never-ma	Adm-cleri	Not-in-far	White	Male	2174	0	40	United-St	<=50K
3	50	Self-emp-	83311	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-St	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-	Not-in-far	White	Male	0	0	40	United-St	<=50K
5	53	Private	234721	11th	7	Married-c	Handlers-	Husband	Black	Male	0	0	40	United-St	<=50K
6	28	Private	338409	Bachelors	13	Married-c	Prof-spec	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St	<=50K
8	49	Private	160187	9th	5	Married-s	Other-ser	Not-in-far	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-	209642	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-St	>50K
10	31	Private	45781	Masters	14	Never-ma	Prof-spec	Not-in-far	White	Female	14084	0	50	United-St	>50K
11	42	Private	159449	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-St	>50K
12	37	Private	280464	Some-coll	10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-St	>50K
13	30	State-gov	141297	Bachelors	13	Married-c	Prof-spec	Husband	Asian-Pac	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-ma	Adm-cleri	Own-chilc	White	Female	0	0	30	United-St	<=50K
15	32	Private	205019	Assoc-acc	12	Never-ma	Sales	Not-in-far	Black	Male	0	0	50	United-St	<=50K
16	40	Private	121772	Assoc-voc	11	Married-c	Craft-repa	Husband	Asian-Pac	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-c	Transport	Husband	Amer-Ind	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-	176756	HS-grad	9	Never-ma	Farming-f	Own-chilc	White	Male	0	0	35	United-St	<=50K
19	32	Private	186824	HS-grad	9	Never-ma	Machine-	Unmarrie	White	Male	0	0	40	United-St	<=50K
20	38	Private	28887	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K
21	43	Self-emp-	292175	Masters	14	Divorced	Exec-man	Unmarrie	White	Female	0	0	45	United-St	>50K



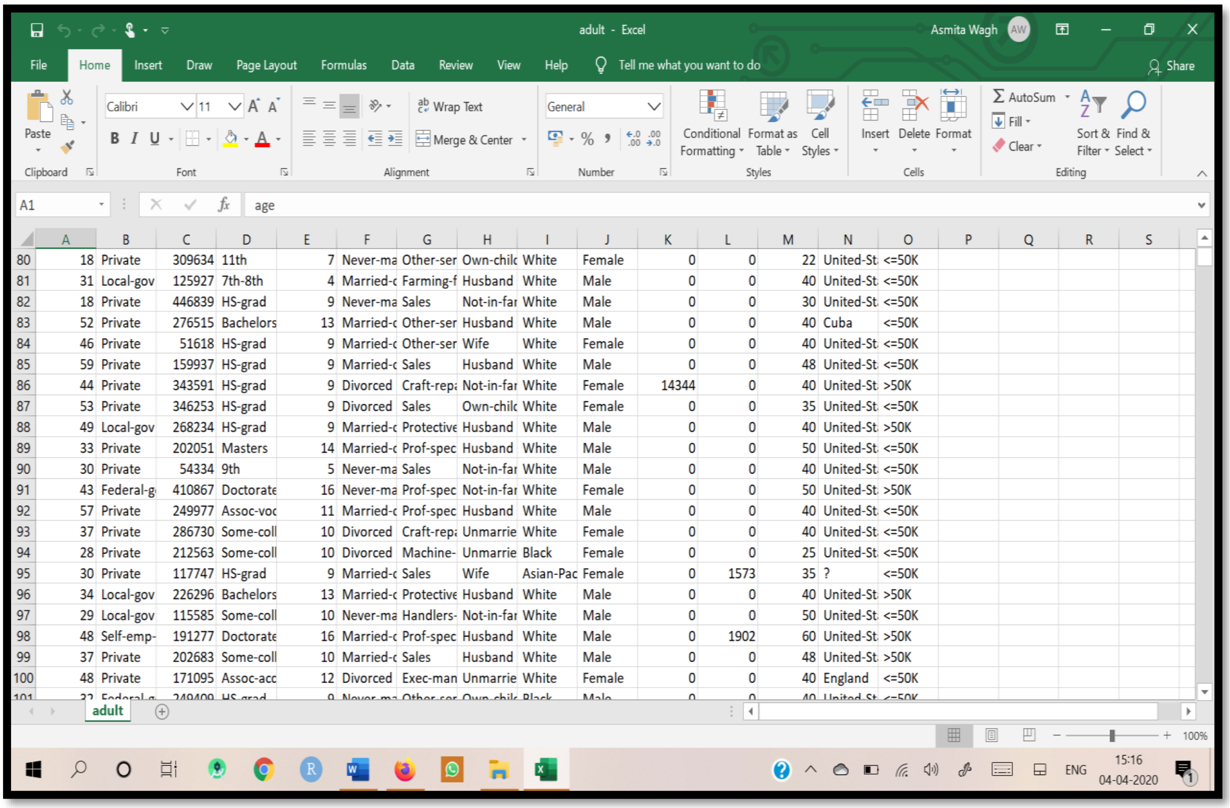
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
20	38	Private	28887	11th	7	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K
21	43	Self-emp-	292175	Masters	14	Divorced	Exec-man	Unmarrie	White	Female	0	0	45	United-St	>50K
22	40	Private	193524	Doctorate	16	Married-c	Prof-spec	Husband	White	Male	0	0	60	United-St	>50K
23	54	Private	302146	HS-grad	9	Separatec	Other-ser	Unmarrie	Black	Female	0	0	20	United-St	<=50K
24	35	Federal-g	76845	9th	5	Married-c	Farming-f	Husband	Black	Male	0	0	40	United-St	<=50K
25	43	Private	117037	11th	7	Married-c	Transport	Husband	White	Male	0	2042	40	United-St	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-sup	Unmarrie	White	Female	0	0	40	United-St	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-c	Tech-sup	Husband	White	Male	0	0	40	United-St	>50K
28	19	Private	168294	HS-grad	9	Never-ma	Craft-repa	Own-chilc	White	Male	0	0	40	United-St	<=50K
29	54	?	180211	Some-coll	10	Married-c	?	Husband	Asian-Pac	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-man	Not-in-far	White	Male	0	0	80	United-St	<=50K
31	49	Private	193366	HS-grad	9	Married-c	Craft-repa	Husband	White	Male	0	0	40	United-St	<=50K
32	23	Local-gov	190709	Assoc-acc	12	Never-ma	Protective	Not-in-far	White	Male	0	0	52	United-St	<=50K
33	20	Private	266015	Some-coll	10	Never-ma	Sales	Own-chilc	Black	Male	0	0	44	United-St	<=50K
34	45	Private	386940	Bachelors	13	Divorced	Exec-man	Own-chilc	White	Male	0	1408	40	United-St	<=50K
35	30	Federal-g	59951	Some-coll	10	Married-c	Adm-cleri	Own-chilc	White	Male	0	0	40	United-St	<=50K
36	22	State-gov	311512	Some-coll	10	Married-c	Other-ser	Husband	Black	Male	0	0	15	United-St	<=50K
37	48	Private	242406	11th	7	Never-ma	Machine-	Unmarrie	White	Male	0	0	40	Puerto-Ri	<=50K
38	21	Private	197200	Some-coll	10	Never-ma	Machine-	Own-chilc	White	Male	0	0	40	United-St	<=50K
39	19	Private	544091	HS-grad	9	Married-A	Adm-cleri	Wife	White	Female	0	0	25	United-St	<=50K
40	31	Private	84154	Some-coll	10	Married-c	Sales	Husband	White	Male	0	0	38	?	>50K



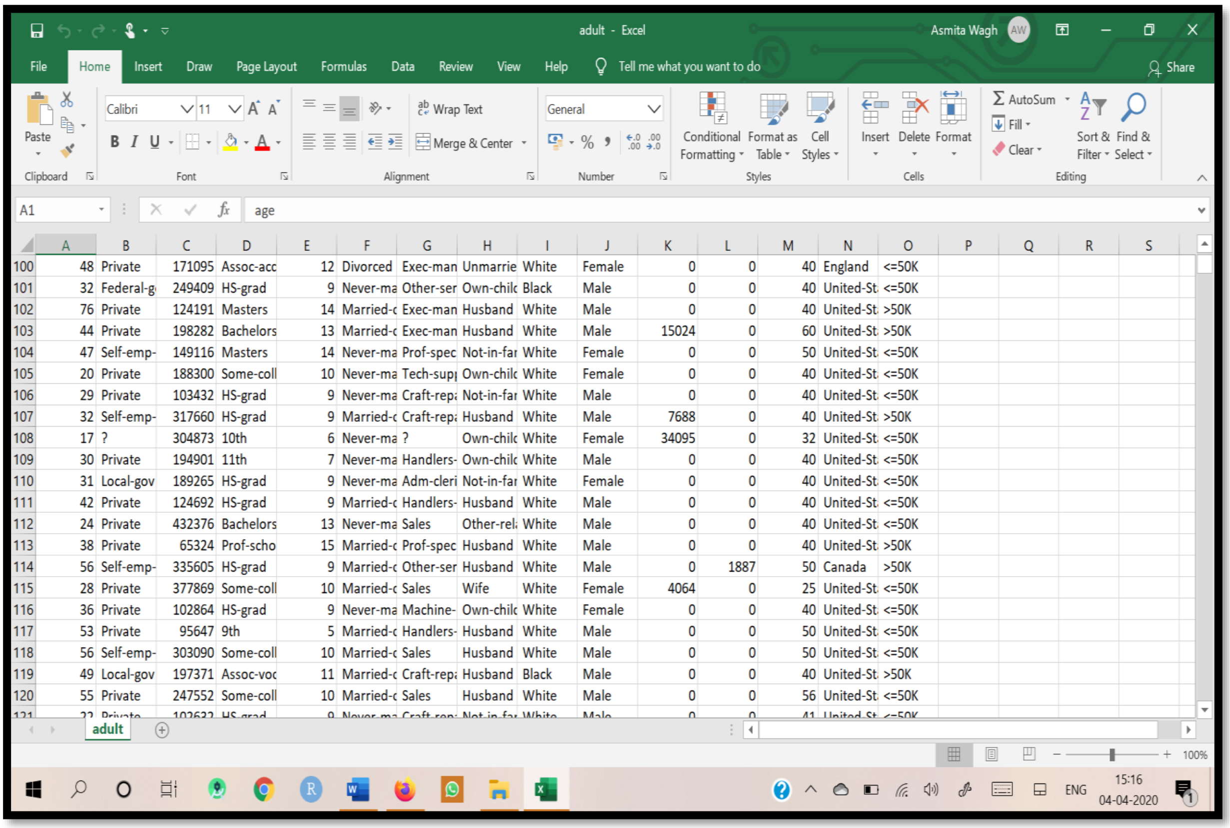
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
41	48	Self-emp-	265477	Assoc-acc	12	Married-c	Prof-spec	Husband	White	Male	0	0	40	United-St	<=50K
42	31	Private	507875	9th	5	Married-c	Machine-	Husband	White	Male	0	0	43	United-St	<=50K
43	53	Self-emp-	88506	Bachelors	13	Married-c	Prof-spec	Husband	White	Male	0	0	40	United-St	<=50K
44	24	Private	172987	Bachelors	13	Married-c	Tech-sup	Husband	White	Male	0	0	50	United-St	<=50K
45	49	Private	94638	HS-grad	9	Separatec	Adm-cleri	Unmarrie	White	Female	0	0	40	United-St	<=50K
46	25	Private	289980	HS-grad	9	Never-ma	Handlers-	Not-in-far	White	Male	0	0	35	United-St	<=50K
47	57	Federal-g	337895	Bachelors	13	Married-c	Prof-spec	Husband	Black	Male	0	0	40	United-St	>50K
48	53	Private	144361	HS-grad	9	Married-c	Machine-	Husband	White	Male	0	0	38	United-St	<=50K
49	44	Private	128354	Masters	14	Divorced	Exec-man	Unmarrie	White	Female	0	0	40	United-St	<=50K
50	41	State-gov	101603	Assoc-voc	11	Married-c	Craft-repa	Husband	White	Male	0	0	40	United-St	<=50K
51	29	Private	271466	Assoc-voc	11	Never-ma	Prof-spec	Not-in-far	White	Male	0	0	43	United-St	<=50K
52	25	Private	32275	Some-coll	10	Married-c	Exec-man	Wife	Other	Female	0	0	40	United-St	<=50K
53	18	Private	226956	HS-grad	9	Never-ma	Other-ser	Own-chilc	White	Female	0	0	30	?	<=50K
54	47	Private	51835	Prof-scho	15	Married-c	Prof-spec	Wife	White	Female	0	1902	60	Honduras	>50K
55	50	Federal-g	251585	Bachelors	13	Divorced	Exec-man	Not-in-far	White	Male	0	0	55	United-St	>50K
56	47	Self-emp-	109832	HS-grad	9	Divorced	Exec-man	Not-in-far	White	Male	0	0	60	United-St	<=50K
57	43	Private	237993	Some-coll	10	Married-c	Tech-sup	Husband	White	Male	0	0	40	United-St	>50K
58	46	Private	216666	5th-6th	3	Married-c	Machine-	Husband	White	Male	0	0	40	Mexico	<=50K
59	35	Private	56352	Assoc-voc	11	Married-c	Other-ser	Husband	White	Male	0	0	40	Puerto-Ri	<=50K
60	41	Private	147372	HS-grad	9	Married-c	Adm-cleri	Husband	White	Male	0	0	48	United-St	<=50K
61	30	Private	188146	HS-grad	9	Married-c	Machine-	Husband	White	Male	5013	0	40	United-St	<=50K
62	30	Private	59496	Bachelors	13	Married-c	Sales	Husband	White	Male	2407	0	40	United-St	<=50K



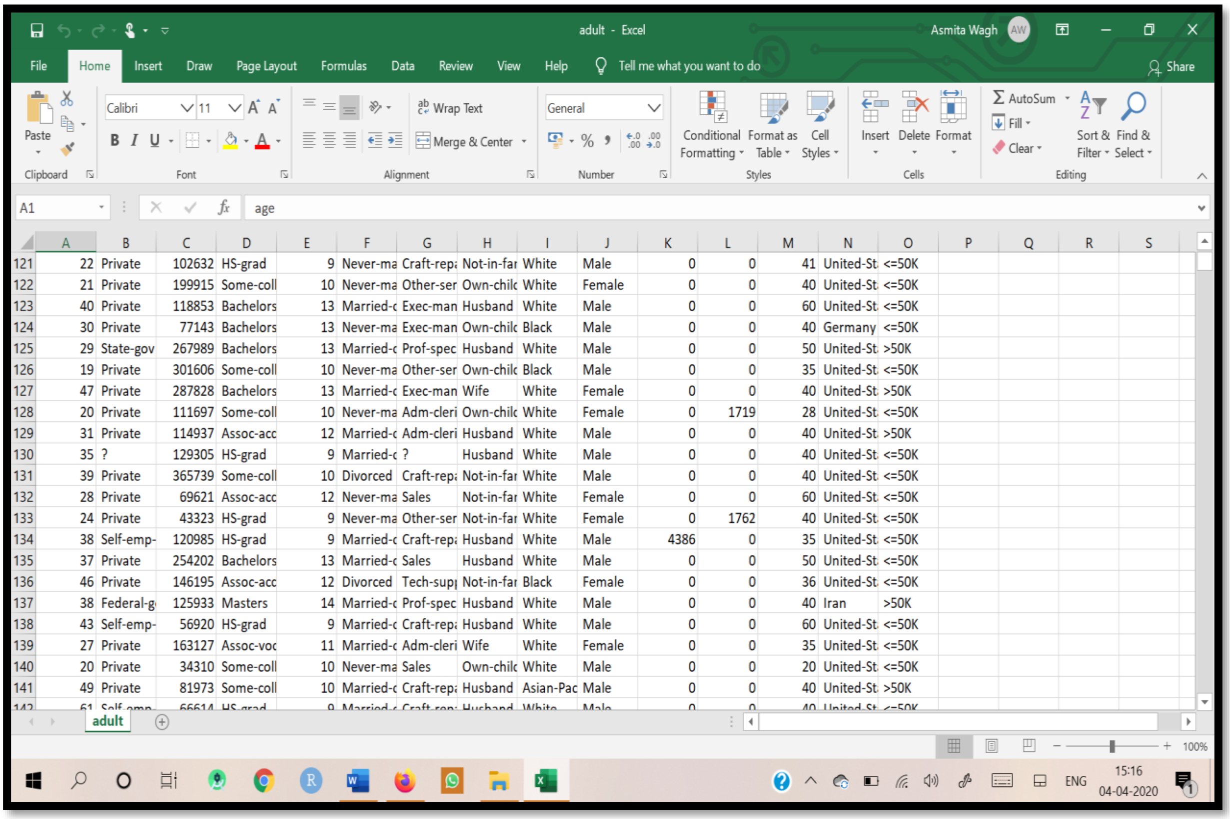
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
60	41	Private	147372	HS-grad	9	Married-c	Adm-cleri	Husband	White	Male	0	0	48	United-St	<=50K
61	30	Private	188146	HS-grad	9	Married-c	Machine-	Husband	White	Male	5013	0	40	United-St	<=50K
62	30	Private	59496	Bachelors	13	Married-c	Sales	Husband	White	Male	2407	0	40	United-St	<=50K
63	32	?	293936	7th-8th	4	Married-s	?	Not-in-far	White	Male	0	0	40	?	<=50K
64	48	Private	149640	HS-grad	9	Married-c	Transport	Husband	White	Male	0	0	40	United-St	<=50K
65	42	Private	116632	Doctorate	16	Married-c	Prof-spec	Husband	White	Male	0	0	45	United-St	>50K
66	29	Private	105598	Some-coll	10	Divorced	Tech-sup	Not-in-far	White	Male	0	0	58	United-St	<=50K
67	36	Private	155537	HS-grad	9	Married-c	Craft-repa	Husband	White	Male	0	0	40	United-St	<=50K
68	28	Private	183175	Some-coll	10	Divorced	Adm-cleri	Not-in-far	White	Female	0	0	40	United-St	<=50K
69	53	Private	169846	HS-grad	9	Married-c	Adm-cleri	Wife	White	Female	0	0	40	United-St	>50K
70	49	Self-emp-	191681	Some-coll	10	Married-c	Exec-man	Husband	White	Male	0	0	50	United-St	>50K
71	25	?	200681	Some-coll	10	Never-ma	?	Own-chilc	White	Male	0	0	40	United-St	<=50K
72	19	Private	101509	Some-coll	10	Never-ma	Prof-spec	Own-chilc	White	Male	0	0	32	United-St	<=50K
73	31	Private	309974	Bachelors	13	Separatec	Sales	Own-chilc	Black	Female	0	0	40	United-St	<=50K
74	29	Self-emp-	162298	Bachelors	13	Married-c	Sales	Husband	White	Male	0	0	70	United-St	>50K
75	23	Private	211678	Some-coll	10	Never-ma	Machine-	Not-in-far	White	Male	0	0	40	United-St	<=50K
76	79	Private	124744	Some-coll	10	Married-c	Prof-spec	Other-rel	White	Male	0	0	20	United-St	<=50K
77	27	Private	213921	HS-grad	9	Never-ma	Other-ser	Own-chilc	White	Male	0	0	40	Mexico	<=50K
78	40	Private	32214	Assoc-acc	12	Married-c	Adm-cleri	Husband	White	Male	0	0	40	United-St	<=50K
79	67	?	212759	10th	6	Married-c	?	Husband	White	Male	0	0	2	United-St	<=50K
80	18	Private	309634	11th	7	Never-ma	Other-ser	Own-chilc	White	Female	0	0	22	United-St	<=50K
81	31	Local-gov	125927	7th-8th	4	Married-c	Farming-f	Husband	White	Male	0	0	40	United-St	<=50K



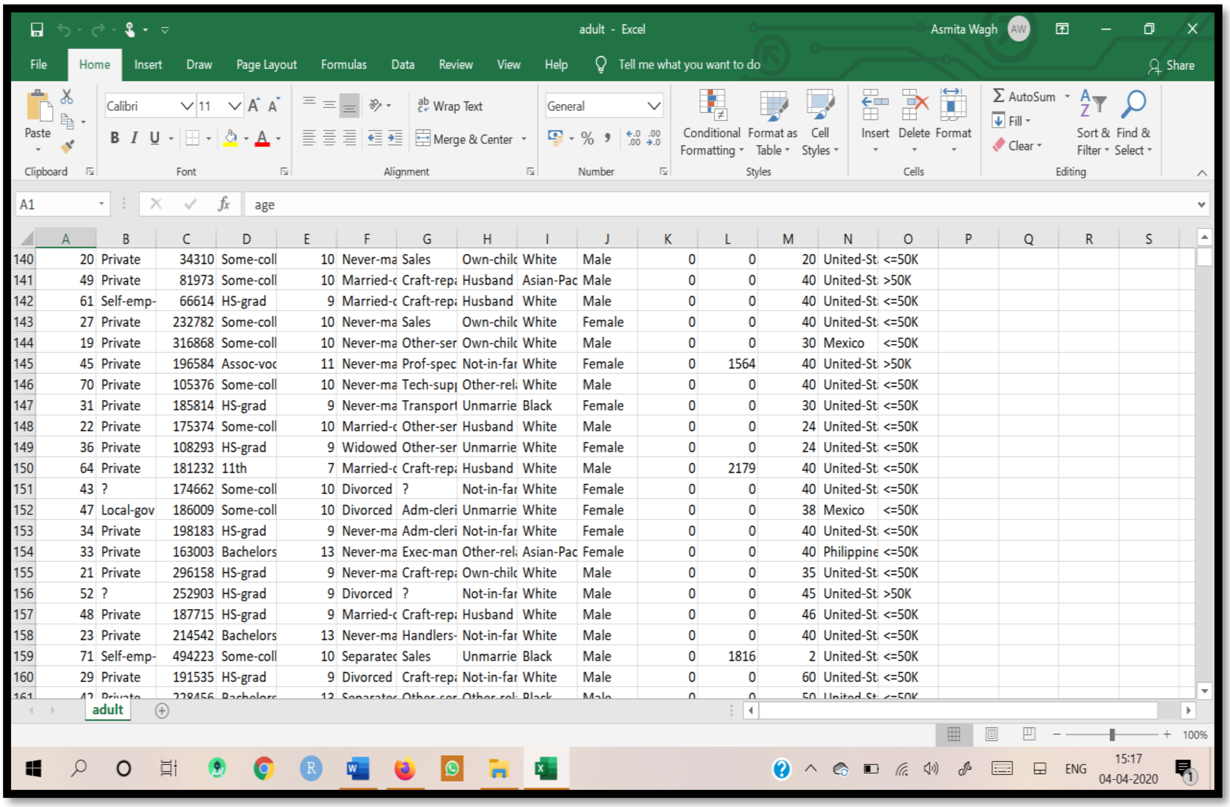
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
80	18	Private	309634	11th	7	Never-ma	Other-ser	Own-chilc	White	Female	0	0	22	United-St	<=50K
81	31	Local-gov	125927	7th-8th	4	Married-c	Farming-f	Husband	White	Male	0	0	40	United-St	<=50K
82	18	Private	446839	HS-grad	9	Never-ma	Sales	Not-in-far	White	Male	0	0	30	United-St	<=50K
83	52	Private	276515	Bachelors	13	Married-c	Other-ser	Husband	White	Male	0	0	40	Cuba	<=50K
84	46	Private	51618	HS-grad	9	Married-c	Other-ser	Wife	White	Female	0	0	40	United-St	<=50K
85	59	Private	159937	HS-grad	9	Married-c	Sales	Husband	White	Male	0	0	48	United-St	<=50K
86	44	Private	343591	HS-grad	9	Divorced	Craft-repa	Not-in-far	White	Female	14344	0	40	United-St	>50K
87	53	Private	346253	HS-grad	9	Divorced	Sales	Own-chilc	White	Female	0	0	35	United-St	<=50K
88	49	Local-gov	268234	HS-grad	9	Married-c	Protective	Husband	White	Male	0	0	40	United-St	>50K
89	33	Private	202051	Masters	14	Married-c	Prof-spec	Husband	White	Male	0	0	50	United-St	<=50K
90	30	Private	54334	9th	5	Never-ma	Sales	Not-in-far	White	Male	0	0	40	United-St	<=50K
91	43	Federal-g	410867	Doctorate	16	Never-ma	Prof-spec	Not-in-far	White	Female	0	0	50	United-St	>50K
92	57	Private	249977	Assoc-voc	11	Married-c	Prof-spec	Husband	White	Male	0	0	40	United-St	<=50K
93	37	Private	286730	Some-coll	10	Divorced	Craft-repa	Unmarrie	White	Female	0	0	40	United-St	<=50K
94	28	Private	212563	Some-coll	10	Divorced	Machine-	Unmarrie	Black	Female	0	0	25	United-St	<=50K
95	30	Private	117747	HS-grad	9	Married-c	Sales	Wife	Asian-Pac	Female	0	1573	35	?	<=50K
96	34	Local-gov	226296	Bachelors	13	Married-c	Protective	Husband	White	Male	0	0	40	United-St	>50K
97	29	Local-gov	115585	Some-coll	10	Never-ma	Handlers-	Not-in-far	White	Male	0	0	50	United-St	<=50K
98	48	Self-emp-	191277	Doctorate	16	Married-c	Prof-spec	Husband	White	Male	0	1902	60	United-St	>50K
99	37	Private	202683	Some-coll	10	Married-c	Sales	Husband	White	Male	0	0	48	United-St	>50K
100	48	Private	171095	Assoc-acc	12	Divorced	Exec-man	Unmarrie	White	Female	0	0	40	England	<=50K



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
100	48	Private	171095	Assoc-acc	12	Divorced	Exec-man	Unmarrie	White	Female	0	0	40	England	<=50K
101	32	Federal-g	249409	HS-grad	9	Never-ma	Other-ser	Own-chilc	Black	Male	0	0	40	United-St	<=50K
102	76	Private	124191	Masters	14	Married-c	Exec-man	Husband	White	Male	0	0	40	United-St	>50K
103	44	Private	198282	Bachelors	13	Married-c	Exec-man	Husband	White	Male	15024	0	60	United-St	>50K
104	47	Self-emp-	149116	Masters	14	Never-ma	Prof-spec	Not-in-far	White	Female	0	0	50	United-St	<=50K
105	20	Private	188300	Some-coll	10	Never-ma	Tech-supp	Own-chilc	White	Female	0	0	40	United-St	<=50K
106	29	Private	103432	HS-grad	9	Never-ma	Craft-repa	Not-in-far	White	Male	0	0	40	United-St	<=50K
107	32	Self-emp-	317660	HS-grad	9	Married-c	Craft-repa	Husband	White	Male	7688	0	40	United-St	>50K
108	17	?	304873	10th	6	Never-ma	?	Own-chilc	White	Female	34095	0	32	United-St	<=50K
109	30	Private	194901	11th	7	Never-ma	Handlers-	Own-chilc	White	Male	0	0	40	United-St	<=50K
110	31	Local-gov	189265	HS-grad	9	Never-ma	Adm-cleri	Not-in-far	White	Female	0	0	40	United-St	<=50K
111	42	Private	124692	HS-grad	9	Married-c	Handlers-	Husband	White	Male	0	0	40	United-St	<=50K
112	24	Private	432376	Bachelors	13	Never-ma	Sales	Other-rel	White	Male	0	0	40	United-St	<=50K
113	38	Private	65324	Prof-scho	15	Married-c	Prof-spec	Husband	White	Male	0	0	40	United-St	>50K
114	56	Self-emp-	335605	HS-grad	9	Married-c	Other-ser	Husband	White	Male	0	1887	50	Canada	>50K
115	28	Private	377869	Some-coll	10	Married-c	Sales	Wife	White	Female	4064	0	25	United-St	<=50K
116	36	Private	102864	HS-grad	9	Never-ma	Machine-	Own-chilc	White	Female	0	0	40	United-St	<=50K
117	53	Private	95647	9th	5	Married-c	Handlers-	Husband	White	Male	0	0	50	United-St	<=50K
118	56	Self-emp-	303090	Some-coll	10	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K
119	49	Local-gov	197371	Assoc-voc	11	Married-c	Craft-repa	Husband	Black	Male	0	0	40	United-St	>50K
120	55	Private	247552	Some-coll	10	Married-c	Sales	Husband	White	Male	0	0	56	United-St	<=50K



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
121	22	Private	102632	HS-grad	9	Never-ma	Craft-rep	Not-in-far	White	Male	0	0	41	United-St	<=50K
122	21	Private	199915	Some-coll	10	Never-ma	Other-ser	Own-chilc	White	Female	0	0	40	United-St	<=50K
123	40	Private	118853	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	60	United-St	<=50K
124	30	Private	77143	Bachelors	13	Never-ma	Exec-man	Own-chilc	Black	Male	0	0	40	Germany	<=50K
125	29	State-gov	267989	Bachelors	13	Married-c	Prof-spec	Husband	White	Male	0	0	50	United-St	>50K
126	19	Private	301606	Some-coll	10	Never-ma	Other-ser	Own-chilc	Black	Male	0	0	35	United-St	<=50K
127	47	Private	287828	Bachelors	13	Married-c	Exec-man	Wife	White	Female	0	0	40	United-St	>50K
128	20	Private	111697	Some-coll	10	Never-ma	Adm-cleri	Own-chilc	White	Female	0	1719	28	United-St	<=50K
129	31	Private	114937	Assoc-acc	12	Married-c	Adm-cleri	Husband	White	Male	0	0	40	United-St	>50K
130	35	?	129305	HS-grad	9	Married-c	?	Husband	White	Male	0	0	40	United-St	<=50K
131	39	Private	365739	Some-coll	10	Divorced	Craft-rep	Not-in-far	White	Male	0	0	40	United-St	<=50K
132	28	Private	69621	Assoc-acc	12	Never-ma	Sales	Not-in-far	White	Female	0	0	60	United-St	<=50K
133	24	Private	43323	HS-grad	9	Never-ma	Other-ser	Not-in-far	White	Female	0	1762	40	United-St	<=50K
134	38	Self-emp-	120985	HS-grad	9	Married-c	Craft-rep	Husband	White	Male	4386	0	35	United-St	<=50K
135	37	Private	254202	Bachelors	13	Married-c	Sales	Husband	White	Male	0	0	50	United-St	<=50K
136	46	Private	146195	Assoc-acc	12	Divorced	Tech-sup	Not-in-far	Black	Female	0	0	36	United-St	<=50K
137	38	Federal-g	125933	Masters	14	Married-c	Prof-spec	Husband	White	Male	0	0	40	Iran	>50K
138	43	Self-emp-	56920	HS-grad	9	Married-c	Craft-rep	Husband	White	Male	0	0	60	United-St	<=50K
139	27	Private	163127	Assoc-voc	11	Married-c	Adm-cleri	Wife	White	Female	0	0	35	United-St	<=50K
140	20	Private	34310	Some-coll	10	Never-ma	Sales	Own-chilc	White	Male	0	0	20	United-St	<=50K
141	49	Private	81973	Some-coll	10	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	United-St	>50K



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
140	20	Private	34310	Some-coll	10	Never-ma	Sales	Own-chilc	White	Male	0	0	20	United-St	<=50K
141	49	Private	81973	Some-coll	10	Married-c	Craft-rep	Husband	Asian-Pac	Male	0	0	40	United-St	>50K
142	61	Self-emp-	66614	HS-grad	9	Married-c	Craft-rep	Husband	White	Male	0	0	40	United-St	<=50K
143	27	Private	232782	Some-coll	10	Never-ma	Sales	Own-chilc	White	Female	0	0	40	United-St	<=50K
144	19	Private	316868	Some-coll	10	Never-ma	Other-ser	Own-chilc	White	Male	0	0	30	Mexico	<=50K
145	45	Private	196584	Assoc-voc	11	Never-ma	Prof-spec	Not-in-far	White	Female	0	1564	40	United-St	>50K
146	70	Private	105376	Some-coll	10	Never-ma	Tech-sup	Other-rel	White	Male	0	0	40	United-St	<=50K
147	31	Private	185814	HS-grad	9	Never-ma	Transport	Unmarrie	Black	Female	0	0	30	United-St	<=50K
148	22	Private	175374	Some-coll	10	Married-c	Other-ser	Husband	White	Male	0	0	24	United-St	<=50K
149	36	Private	108293	HS-grad	9	Widowed	Other-ser	Unmarrie	White	Female	0	0	24	United-St	<=50K
150	64	Private	181232	11th	7	Married-c	Craft-rep	Husband	White	Male	0	2179	40	United-St	<=50K
151	43	?	174662	Some-coll	10	Divorced	?	Not-in-far	White	Female	0	0	40	United-St	<=50K
152	47	Local-gov	186009	Some-coll	10	Divorced	Adm-cleri	Unmarrie	White	Female	0	0	38	Mexico	<=50K
153	34	Private	198183	HS-grad	9	Never-ma	Adm-cleri	Not-in-far	White	Female	0	0	40	United-St	<=50K
154	33	Private	163003	Bachelors	13	Never-ma	Exec-man	Other-rel	Asian-Pac	Female	0	0	40	Philippine	<=50K
155	21	Private	296158	HS-grad	9	Never-ma	Craft-rep	Own-chilc	White	Male	0	0	35	United-St	<=50K
156	52	?	252903	HS-grad	9	Divorced	?	Not-in-far	White	Male	0	0	45	United-St	>50K
157	48	Private	187715	HS-grad	9	Married-c	Craft-rep	Husband	White	Male	0	0	46	United-St	<=50K
158	23	Private	214542	Bachelors	13	Never-ma	Handlers-	Not-in-far	White	Male	0	0	40	United-St	<=50K
159	71	Self-emp-	494223	Some-coll	10	Separatec	Sales	Unmarrie	Black	Male	0	1816	2	United-St	<=50K
160	29	Private	191535	HS-grad	9	Divorced	Craft-rep	Not-in-far	White	Male	0	0	60	United-St	<=50K

Theory:

The data was extracted from 1994 Census bureau database. Please view the adult.csv file for the dataset, this dataset is what is being used in the entire study of ours.

income <- read.csv ('adult.csv', na. strings = c ('‘, ‘?’)) ……… To assign the csv data into variable income, the na. strings are basically used to specify the content that is unknown in the dataset or left blank.

supply (income, function(x) sum(is.na(x))) ………. basically, counts the data that is unavailable.

After finding it out, I realized that numbers are a bit annoying to comprehend and select the reliable points, henceforth we went ahead and plotted the graphs of it to see how they are.

```
library(Amelia)
missmap(income, main = "Missing values vs observed")
table (complete.cases (income))`
```

This basically allows us to see the plot of missing and observed data.

Check Rplots.pdf for the plot

Since the dataset has already been cleaned running it up again, will not show any false values.

```
library(Amelia)
Loading required package: foreign
##
## Amelia II: Multiple Imputation
## (Version 1.6.4, built: 2012-12-17)
## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell
## Refer to http://gking.harvard.edu/amelia/ for more information
##
> missmap(income, main = "Missing values vs observed")
> table (complete.cases (income))

 TRUE
32561
```

So, we go ahead and make a boxplot of everything with respect to income level.

```
library(gridExtra)
p1 <- ggplot(aes(x=income, y=age), data = income) + geom_boxplot() +
  ggtitle('Age vs. Income Level')
p2 <- ggplot(aes(x=income, y=education.num), data = income) + geom_boxplot() +
  ggtitle('Years of Education vs. Income Level')
str(income)
p3 <- ggplot(aes(x=income, y=house.per.week), data = income) + geom_boxplot() + ggtitle('Hours Per week vs. Income Level
p4 <- ggplot(aes(x=income, y=capital.gain), data=income) + geom_boxplot() +
  ggtitle('Capital Gain vs. Income Level')
p5 <- ggplot(aes(x=income, y=capital.loss), data=income) + geom_boxplot() +
  ggtitle('Capital Loss vs. Income Level')
p6 <- ggplot(aes(x=income, y=fnlwgt), data=income) + geom_boxplot() +
  ggtitle('Final Weight vs. Income Level')
grid.arrange(p1, p2, p3, p4, p5, p6, ncol=3)
income$fnlwgt <- NULL
```

“Age”, “Years of education” and “hours per week” all show significant variations with income level. Therefore, they are kept for the regression analysis. “Final Weight” does not show any variation with income level, therefore, it has been excluded from the analysis. It’s hard to see whether “Capital gain” and “Capital loss” have variation with Income level from the above plot, so we shall keep them for now.

```
library(dplyr)
by_workclass <- income %>% group_by(workclass, income) %>% summarise(n=n())
by_education <- income %>% group_by(education, income) %>% summarise(n=n())
by_education$education <- ordered(by_education$education,
                                   levels = c('Preschool', '1st-4th', '5th-6th', '7th-8th', '9th', '10th', '11th', '12th
by_marital <- income %>% group_by(marital.status, income) %>% summarise(n=n())
by_occupation <- income %>% group_by(occupation, income) %>% summarise(n=n())
by_relationship <- income %>% group_by(relationship, income) %>% summarise(n=n())
by_race <- income %>% group_by(race, income) %>% summarise(n=n())
by_sex <- income %>% group_by(sex, income) %>% summarise(n=n())
by_country <- income %>% group_by(native.country, income) %>% summarise(n=n())
p7 <- ggplot(aes(x=workclass, y=n, fill=income), data=by_workclass) + geom_bar(stat = 'identity', position = position_do
p8 <- ggplot(aes(x=education, y=n, fill=income), data=by_education) + geom_bar(stat = 'identity', position = position_do
p9 <- ggplot(aes(x=marital.status, y=n, fill=income), data=by_marital) + geom_bar(stat = 'identity', position=position_d
p10 <- ggplot(aes(x=occupation, y=n, fill=income), data=by_occupation) + geom_bar(stat = 'identity', position=position_d
p11 <- ggplot(aes(x=relationship, y=n, fill=income), data=by_relationship) + geom_bar(stat = 'identity', position=positi
p12 <- ggplot(aes(x=race, y=n, fill=income), data=by_race) + geom_bar(stat = 'identity', position = position_dodge()) +
p13 <- ggplot(aes(x=sex, y=n, fill=income), data=by_sex) + geom_bar(stat = 'identity', position = position_dodge()) + gg
p14 <- ggplot(aes(x=native.country, y=n, fill=income), data=by_country) + geom_bar(stat = 'identity', position = positio
grid.arrange(p7, p8, p9, p10, ncol=2)
grid.arrange(p11,p12,p13, ncol=2)
```

Most of the data is collected from the United States, so variable “native country” does not have effect on our analysis, we shall exclude it from regression model.

And all the other categorial variables seem to have reasonable variation, so they will be kept.

income$income = as.factor(ifelse(income$income==income$income[1],0,1)) basically if >50k it will be set to 1 else to 0

```r
train <- income[1:22793,]
test <- income[22793:32561,]
model <-glm(income ~.,family=binomial(link='logit'),data=train)
summary(model)
```

So as to fit the model and view the details.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9275  -0.5171  -0.1885  -0.0327   3.2694

Coefficients: (2 not defined because of singularities)
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -8.325e+00  4.908e-01 -16.963  < 2e-16 ***
age                                   2.505e-02  1.932e-03  12.961  < 2e-16 ***
workclass Federal-gov                 1.100e+00  1.818e-01   6.055 1.41e-09 ***
workclass Local-gov                   4.509e-01  1.660e-01   2.716 0.006599 **
workclass Never-worked               -9.225e+00  6.808e+02  -0.014 0.989189
workclass Private                     6.245e-01  1.481e-01   4.216 2.49e-05 ***
workclass Self-emp-inc                8.046e-01  1.766e-01   4.557 5.20e-06 ***
workclass Self-emp-not-inc            1.674e-01  1.623e-01   1.031 0.302392
workclass State-gov                   2.851e-01  1.804e-01   1.580 0.114139
workclass Without-pay                -1.279e+01  4.165e+02  -0.031 0.975497
education 11th                        3.831e-02  2.421e-01   0.158 0.874287
education 12th                        3.946e-01  3.182e-01   1.240 0.214856
education 1st-4th                    -5.194e-01  5.324e-01  -0.976 0.329244
education 5th-6th                    -4.645e-01  3.784e-01  -1.228 0.219579
education 7th-8th                    -4.160e-01  2.683e-01  -1.550 0.121113
education 9th                        -4.346e-01  3.154e-01  -1.378 0.168233
education Assoc-acdm                  1.312e+00  2.054e-01   6.388 1.68e-10 ***
education Assoc-voc                   1.307e+00  1.972e-01   6.627 3.43e-11 ***
education Bachelors                   1.885e+00  1.821e-01  10.351  < 2e-16 ***
education Doctorate                   2.840e+00  2.522e-01  11.261  < 2e-16 ***
education HS-grad                     7.689e-01  1.773e-01   4.336 1.45e-05 ***
education Masters                     2.181e+00  1.947e-01  11.204  < 2e-16 ***
```

```
education Preschool                  -1.996e+01  2.167e+02  -0.092 0.926598
education Prof-school                 2.614e+00  2.353e-01  11.109  < 2e-16 ***
education Some-college                1.062e+00  1.802e-01   5.897 3.71e-09 ***
education.num                                NA         NA      NA       NA
marital.status Married-AF-spouse      2.687e+00  6.768e-01   3.970 7.19e-05 ***
marital.status Married-civ-spouse     2.246e+00  3.188e-01   7.043 1.88e-12 ***
marital.status Married-spouse-absent -2.308e-02  2.629e-01  -0.088 0.930043
marital.status Never-married         -3.864e-01  1.039e-01  -3.718 0.000201 ***
marital.status Separated              5.553e-02  1.931e-01   0.288 0.773704
marital.status Widowed                1.271e-01  1.917e-01   0.663 0.507395
occupation Adm-clerical               1.184e-02  1.163e-01   0.102 0.918946
occupation Armed-Forces              -1.325e+01  5.210e+02  -0.025 0.979717
occupation Craft-repair               1.318e-01  9.983e-02   1.320 0.186672
occupation Exec-managerial            8.385e-01  1.027e-01   8.166 3.19e-16 ***
occupation Farming-fishing           -9.949e-01  1.696e-01  -5.867 4.45e-09 ***
occupation Handlers-cleaners         -6.268e-01  1.732e-01  -3.620 0.000295 ***
occupation Machine-op-inspct         -3.041e-01  1.256e-01  -2.420 0.015510 *
occupation Other-service             -7.233e-01  1.435e-01  -5.041 4.63e-07 ***
occupation Priv-house-serv           -1.205e+01  1.174e+02  -0.103 0.918289
occupation Prof-specialty             6.148e-01  1.103e-01   5.573 2.50e-08 ***
occupation Protective-serv            5.988e-01  1.526e-01   3.924 8.70e-05 ***
occupation Sales                      2.675e-01  1.058e-01   2.527 0.011507 *
occupation Tech-support               7.188e-01  1.403e-01   5.125 2.98e-07 ***
occupation Transport-moving                  NA         NA      NA       NA
relationship Not-in-family            5.715e-01  3.148e-01   1.815 0.069470 .
relationship Other-relative          -5.502e-01  2.934e-01  -1.876 0.060718 .
relationship Own-child               -6.178e-01  3.113e-01  -1.984 0.047209 *
relationship Unmarried                3.774e-01  3.342e-01   1.129 0.258751
relationship Wife                     1.357e+00  1.213e-01  11.193  < 2e-16 ***
```

```
relationship Wife                  1.357e+00  1.213e-01  11.193  < 2e-16 ***
race Asian-Pac-Islander            2.078e-01  2.824e-01   0.736 0.461944
race Black                         3.286e-01  2.690e-01   1.222 0.221864
race Other                        -5.340e-01  4.306e-01  -1.240 0.214975
race White                         4.402e-01  2.565e-01   1.716 0.086104 .
sex Male                           8.320e-01  9.420e-02   8.832  < 2e-16 ***
capital.gain                       3.039e-04  1.187e-05  25.598  < 2e-16 ***
capital.loss                       6.484e-04  4.445e-05  14.587  < 2e-16 ***
house.per.week                     2.919e-02  1.934e-03  15.094  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25044  on 22792  degrees of freedom
Residual deviance: 14581  on 22736  degrees of freedom
AIC: 14695

Number of Fisher Scoring iterations: 14

Analysis of Deviance Table

Model: binomial, link: logit

Response: income

Terms added sequentially (first to last)
```

```
               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           22792      25044
age             1   1152.1     22791      23892 < 2.2e-16 ***
workclass       8    558.3     22783      23334 < 2.2e-16 ***
education      15   2551.2     22768      20783 < 2.2e-16 ***
education.num   0      0.0     22768      20783
marital.status  6   3687.4     22762      17096 < 2.2e-16 ***
occupation     13    552.5     22749      16543 < 2.2e-16 ***
relationship    5    165.6     22744      16377 < 2.2e-16 ***
race            4     19.9     22740      16358 0.0005324 ***
sex             1    105.8     22739      16252 < 2.2e-16 ***
capital.gain    1   1210.5     22738      15041 < 2.2e-16 ***
capital.loss    1    227.1     22737      14814 < 2.2e-16 ***
house.per.week  1    233.4     22736      14581 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "Accuracy : 0.851776026205343"
```

Which gives an accuracy of upto 85%.

Code:

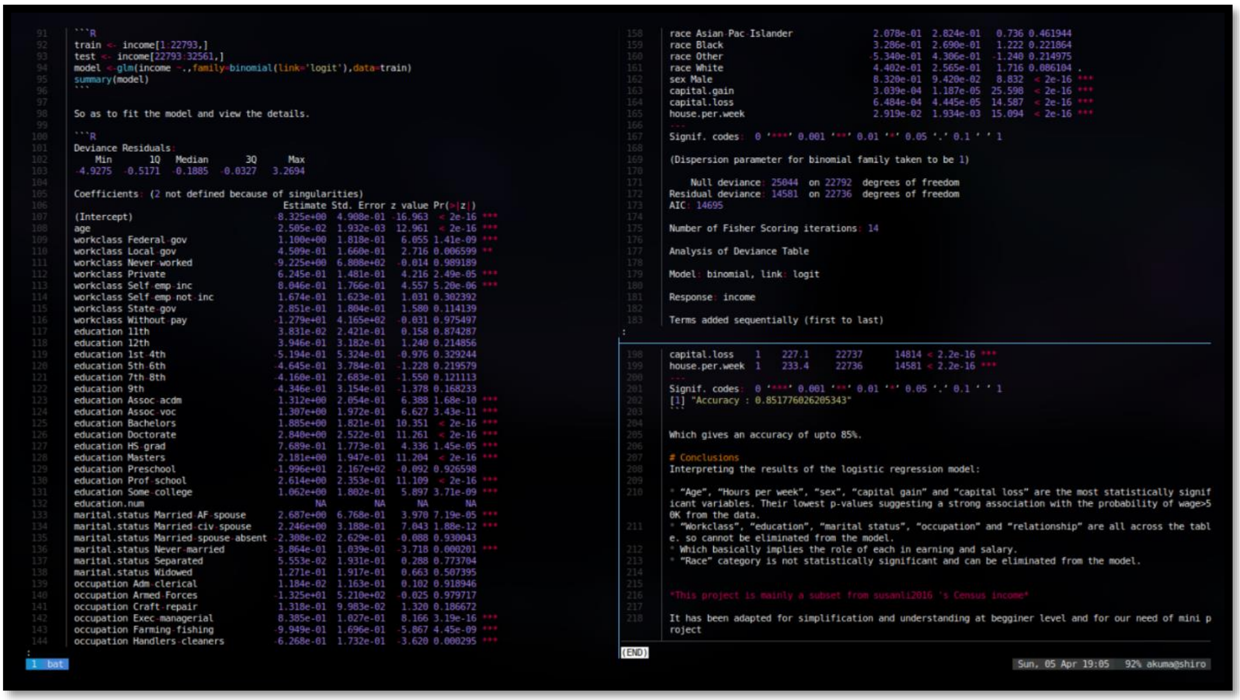
```
q()
income <- read.csv('adult.csv',na.strings = c('','?'))
```

```
str(income0
str(income)
summary(income)
sapply(income,function(x) sum(is.na(x)))
sapply(income, function(x) length(unique(x)))
library(Amelia)
missmap(income, main = "Missing values vs observed")
table (complete.cases (income))
income <- income[complete.cases(income),]
library(ggplot2)
library(gridExtra)
p1 <- ggplot(aes(x=income, y=age), data = income) + geom_boxplot() +
    ggtitle('Age vs. Income Level')
p2 <- ggplot(aes(x=income, y=education.num), data = income) +
geom_boxplot() +
ggtitle('Years of Education vs. Income Level')
str(incomeo)
str(income)
p3 <- ggplot(aes(x=income, y=hours.per.week), data = income) +
geom_boxplot() +
p3 <- ggplot(aes(x=income, y=house.per.week), data = income) +
geom_boxplot() +
p3 <- ggplot(aes(x=income, y=house.per.week), data = income) +
geom_boxplot() + ggtitle('Hours Per week vs. Income Level')
p4 <- ggplot(aes(x=income, y=capital.gain), data=income) +
geom_boxplot()+
ggtitle('Capital Gain vs. Income Level')
p5 <- ggplot(aes(x=income, y=capital.loss), data=income) +
geom_boxplot()+
ggtitle('Capital Loss vs. Income Level')
p6 <- ggplot(aes(x=income, y=fnlwgt), data=income) + geom_boxplot() +
ggtitle('Final Weight vs. Income Level')
grid.arrange(p1, p2, p3, p4, p5, p6, ncol=3)
income$fnlwgt <- NULL
#cuz final weight shows no variation wrt income
library(dplyr)
by_workclass <- income %>% group_by(workclass, income) %>%
summarise(n=n())
by_education <- income %>% group_by(education, income) %>%
summarise(n=n())
```

```
by_education$education <- ordered(by_education$education,
levels = c('Preschool', '1st-4th', '5th-6th', '7th-8th', '9th', '10th', '11th', '12th', 'HS-grad', 'Prof-school', 'Assoc-acdm', 'Assoc-voc', 'Some-college', 'Bachelors', 'Masters', 'Doctorate'))
by_marital <- income %>% group_by(marital.status, income) %>% summarise(n=n())
by_occupation <- income %>% group_by(occupation, income) %>% summarise(n=n())
by_relationship <- income %>% group_by(relationship, income) %>% summarise(n=n())
by_race <- income %>% group_by(race, income) %>% summarise(n=n())
by_sex <- income %>% group_by(sex, income) %>% summarise(n=n())
by_country <- income %>% group_by(native.country, income) %>% summarise(n=n())
p7 <- ggplot(aes(x=workclass, y=n, fill=income), data=by_workclass) + geom_bar(stat = 'identity', position = position_dodge()) + ggtitle('Workclass with Income Level') + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p8 <- ggplot(aes(x=education, y=n, fill=income), data=by_education) + geom_bar(stat = 'identity', position = position_dodge()) + ggtitle('Education vs. Income Level') + coord_flip()
p9 <- ggplot(aes(x=marital.status, y=n, fill=income), data=by_marital) + geom_bar(stat = 'identity', position=position_dodge()) + ggtitle('Marital Status vs. Income Level') + theme(axis.text.x = element_text(angle = 45, hjust = 1))
p10 <- ggplot(aes(x=occupation, y=n, fill=income), data=by_occupation) + geom_bar(stat = 'identity', position=position_dodge()) + ggtitle('Occupation vs. Income Level') + coord_flip()
p11 <- ggplot(aes(x=relationship, y=n, fill=income), data=by_relationship) + geom_bar(stat = 'identity', position=position_dodge()) + ggtitle('Relationship vs. Income Level') + coord_flip()
p12 <- ggplot(aes(x=race, y=n, fill=income), data=by_race) + geom_bar(stat = 'identity', position = position_dodge()) + ggtitle('Race vs. Income Level') + coord_flip()
p13 <- ggplot(aes(x=sex, y=n, fill=income), data=by_sex) + geom_bar(stat = 'identity', position = position_dodge()) + ggtitle('Sex vs. Income Level')
p14 <- ggplot(aes(x=native.country, y=n, fill=income), data=by_country) + geom_bar(stat = 'identity', position = position_dodge()) + ggtitle('Native Country vs. Income Level') + coord_flip()
grid.arrange(p7, p8, p9, p10, ncol=2)
```

```
#categorical variable exploration and plotting
income$native.country <- NULL
#for simplification and saving myself from headache I prefer to not conduct my study as per the countries
income$income = as.factor(ifelse(income$income==income$income[1],0,1))
#basically if >50k it will be set to 1 else to 0
summary(income)
str(income)
len(income)
count(income)
train <- income(22793)
train <- income[1:22793,]
test <- income[22793:32561,]
model <-glm(income ~.,family=binomial(link='logit'),data=train)
summary(model)
anova(model,test="Chisq")
fitted.results <- predict(model,newdata=test,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != test$income)
print(paste('Accuracy :',1-misClasificError))
library(ROCR)
install.packages(ROCR)
R CMD INSTALL ROCR_1.0-1.tar.gz
install.packages(gplot)
q()
getwd()
ls()
savehistory(file="SalaryPrediction")
```

Output:



```
```R
train <- income[1:22793,]
test <- income[22793:32561,]
model <-glm(income ~.,family=binomial(link='logit'),data=train)
summary(model)
```

So as to fit the model and view the details.

```R
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9275  -0.5171  -0.1885  -0.0327   3.2694

Coefficients: (2 not defined because of singularities)
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -8.325e+00  4.908e-01 -16.963  < 2e-16 ***
age                                 2.505e-02  1.932e-03  12.961  < 2e-16 ***
workclass Federal-gov               1.100e+00  1.818e-01   6.055 1.41e-09 ***
workclass Local-gov                 4.509e-01  1.660e-01   2.716 0.006599 **
workclass Never-worked             -9.225e+00  6.808e+02  -0.014 0.989189
workclass Private                   6.245e-01  1.481e-01   4.216 2.49e-05 ***
workclass Self-emp-inc              8.046e-01  1.766e-01   4.557 5.20e-06 ***
workclass Self-emp-not-inc          1.674e-01  1.623e-01   1.031 0.302392
workclass State-gov                 2.851e-01  1.804e-01   1.580 0.114139
workclass Without-pay              -1.279e+01  4.165e+02  -0.031 0.975497
education 11th                      3.831e-02  2.421e-01   0.158 0.874287
education 12th                      3.946e-01  3.182e-01   1.240 0.214856
education 1st-4th                  -5.194e-01  5.324e-01  -0.976 0.329244
education 5th-6th                  -4.645e-01  3.784e-01  -1.228 0.219579
education 7th-8th                  -4.160e-01  2.683e-01  -1.550 0.121113
education 9th                      -4.346e-01  3.154e-01  -1.378 0.168233
education Assoc-acdm                1.312e+00  2.054e-01   6.388 1.68e-10 ***
education Assoc-voc                 1.307e+00  1.972e-01   6.627 3.43e-11 ***
education Bachelors                 1.885e+00  1.821e-01  10.351  < 2e-16 ***
education Doctorate                 2.840e+00  2.522e-01  11.261  < 2e-16 ***
education HS-grad                   7.689e-01  1.773e-01   4.336 1.45e-05 ***
education Masters                   2.181e+00  1.947e-01  11.204  < 2e-16 ***
education Preschool                -1.996e+01  2.167e+02  -0.092 0.926598
education Prof-school               2.614e+00  2.353e-01  11.109  < 2e-16 ***
education Some-college              1.062e+00  1.802e-01   5.897 3.71e-09 ***
education.num                              NA         NA      NA       NA
marital.status Married-AF-spouse    2.687e+00  6.768e-01   3.970 7.19e-05 ***
marital.status Married-civ-spouse   2.246e+00  3.188e-01   7.043 1.88e-12 ***
marital.status Married-spouse-absent -2.308e-02 2.629e-01  -0.088 0.930043
marital.status Never-married       -3.864e-01  1.039e-01  -3.718 0.000201 ***
marital.status Separated            5.553e-02  1.931e-01   0.288 0.773704
marital.status Widowed              1.271e-01  1.917e-01   0.663 0.507395
occupation Adm-clerical             1.184e-02  1.163e-01   0.102 0.918946
occupation Armed-Forces            -1.325e+01  5.210e+02  -0.025 0.979717
occupation Craft-repair             1.318e-01  9.983e-02   1.320 0.186672
occupation Exec-managerial          8.385e-01  1.027e-01   8.166 3.19e-16 ***
occupation Farming-fishing         -9.949e-01  1.696e-01  -5.867 4.45e-09 ***
occupation Handlers-cleaners       -6.268e-01  1.732e-01  -3.620 0.000295 ***
:
 1  bat

race Asian-Pac-Islander             2.078e-01  2.824e-01   0.736 0.461944
race Black                          3.286e-01  2.690e-01   1.222 0.221864
race Other                         -5.340e-01  4.306e-01  -1.240 0.214975
race White                          4.402e-01  2.565e-01   1.716 0.086104 .
sex Male                            8.320e-01  9.420e-02   8.832  < 2e-16 ***
capital.gain                        3.039e-04  1.187e-05  25.598  < 2e-16 ***
capital.loss                        6.484e-04  4.445e-05  14.587  < 2e-16 ***
house.per.week                      2.919e-02  1.934e-03  15.094  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25044  on 22792  degrees of freedom
Residual deviance: 14581  on 22736  degrees of freedom
AIC: 14695

Number of Fisher Scoring iterations: 14

Analysis of Deviance Table

Model: binomial, link: logit

Response: income

Terms added sequentially (first to last)
:
capital.loss    1    227.1     22737      14814 < 2.2e-16 ***
house.per.week  1    233.4     22736      14581 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "Accuracy : 0.851776026205343"
```

Which gives an accuracy of upto 85%.

Conclusions
Interpreting the results of the logistic regression model:

* "Age", "Hours per week", "sex", "capital gain" and "capital loss" are the most statistically signif
icant variables. Their lowest p-values suggesting a strong association with the probability of wage>5
0K from the data.
* "Workclass", "education", "marital status", "occupation" and "relationship" are all across the tabl
e. so cannot be eliminated from the model.
* Which basically implies the role of each in earning and salary.
* "Race" category is not statistically significant and can be eliminated from the model.

This project is mainly a subset from susanli2016 's Census income

It has been adapted for simplification and understanding at begginer level and for our need of mini p
roject
(END)
Sun, 05 Apr 19:05   92% akuma@shiro
```

Conclusion:

Interpreting the results of the logistic regression model:

- "Age", "Hours per week", "sex", "capital gain" and "capital loss" are the most statistically significant variables. Their lowest p-values suggesting a strong association with the probability of wage>50K from the data.
- "Workclass", "education", "marital status", "occupation" and "relationship" are all across the table. so, cannot be eliminated from the model.
- Which basically implies the role of each in earning and salary.
- "Race" category is not statistically significant and can be eliminated from the model.