Machine Learning for Problem Solving(Spring 2018)

Anupam Dewan(adewan)

Question 5: Exploraratory Data Analysis and KDE

5.1 Summary Statistics and what are meaningless qunantities?

EmployeeID	recorddate_key	birthdate_key	orighiredate_key	terminationdate_key	age	length_of_service	city_name	department_name	job_title	s
Min. :1318	12/31/2013 0:00: 5215	1954-08-04: 40	1992-08-09: 50	1900-01-01:42450	Min. :19.00	Min. : 0.00	Vancouver :11211	Meats :10269	Meat Cutter :9984	
1st Qu.:3360	12/31/2012 0:00: 5101	1956-04-27: 40	1995-02-22: 50	2014-12-30: 1079	1st Qu.:31.00	1st Qu.: 5.00	Victoria : 4885	Dairy : 8599	Dairy Person :8590	
Median :5031	12/31/2011 0:00: 4972	1973-03-23: 40	2004-12-04: 50	2015-12-30: 674	Median :42.00	Median :10.00	Nanaimo : 3876	Produce : 8515	Produce Clerk:8237	
Mean :4859	12/31/2014 0:00: 4962	1952-01-27: 30	2005-10-16: 50	2010-12-30: 25	Mean :42.08	Mean :10.43	New Westminster: 3211	Bakery : 8381	Baker :8096	M
3rd Qu.:6335	12/31/2010 0:00: 4840	1952-08-10: 30	2006-02-26: 50	2012-11-11: 21	3rd Qu.:53.00	3rd Qu.:15.00	Kelowna : 2513	Customer Service: 7122	Cashier :6816	
Max. :8336	12/31/2015 0:00: 4799	1953-10-06: 30	2006-09-25: 50	2015-02-04: 20	Max. :65.00	Max. :26.00	Burnaby : 2067	Processed Foods : 5911	Shelf Stocker:5622	
NA	(Other) :19764	(Other) :49443	(Other) :49353	(Other) : 5384	NA	NA	(Other) :21890	(Other) : 856	(Other) :2308	

According to summary statistics it seems that Gender_full and Gender_Short represent the same thing (one says “M” and Males which essentially means the same) and therefore we eleminate one of them.(eliminated Gender_full)

5.2 The percentage of terminated employees out of all employees for each year and Average termination rate in 10 years

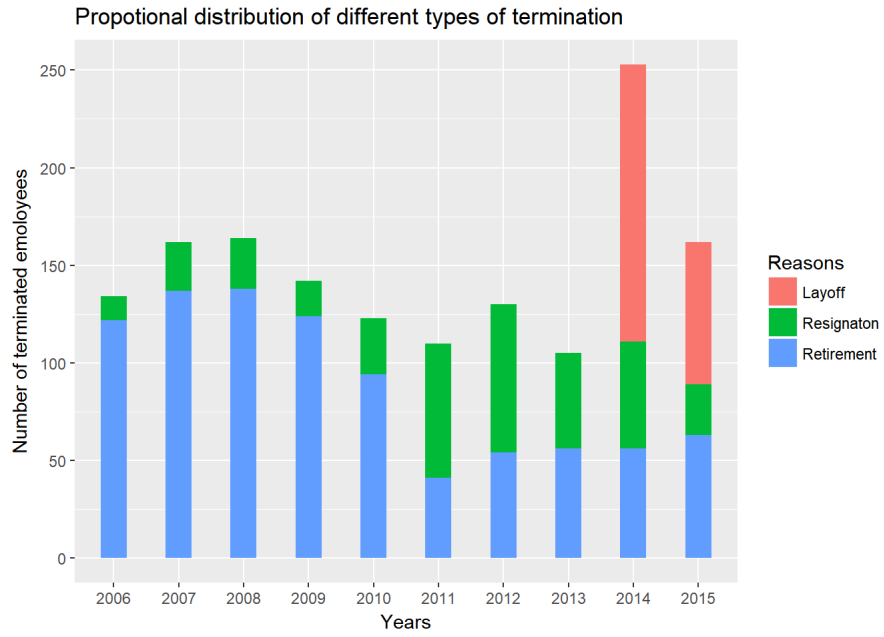
STATUS_YEAR	percentage
2006	2.926403
2007	3.459321
2008	3.440319
2009	2.926628
2010	2.478340
2011	2.164502

STATUS_YEAR	percentage
2012	2.485184
2013	1.973684
2014	4.851390
2015	3.265471

2.997124

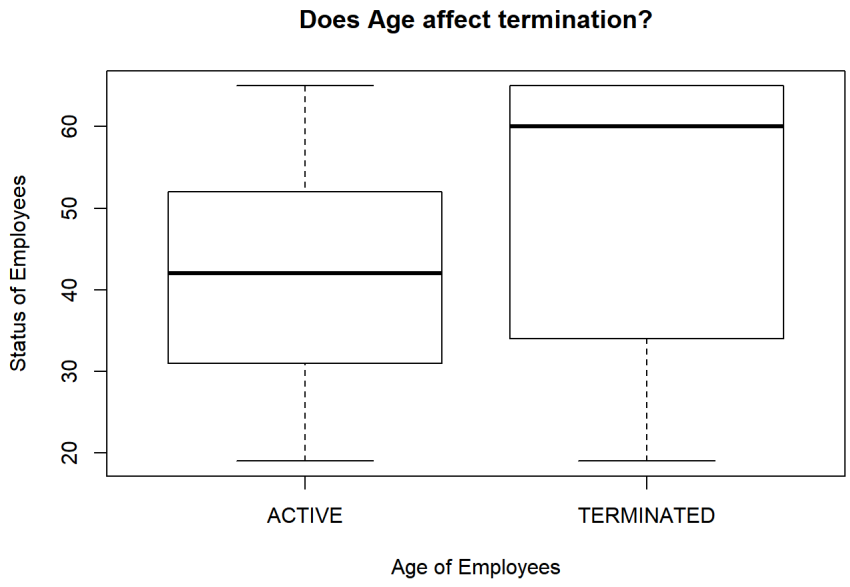
As can be seen from table above the percentage of termination seems to be in the rangeof 2-3% year on year basis. Moreover, the average termination rate is around 3%

5.3 A stacked bar chart of terminates.



Interpretation: Majorly from 2006-2010 there was no layoff per se, and the major contribution of termination of employee service was due to the retirement. From 2010-2013 the distribution started to spread out evenly and almost similar percentage of employees were terminating due to either of the two reasons. Post 2013, company has started to lay off its employees which has then become one of the major reasons for employee service terminations.

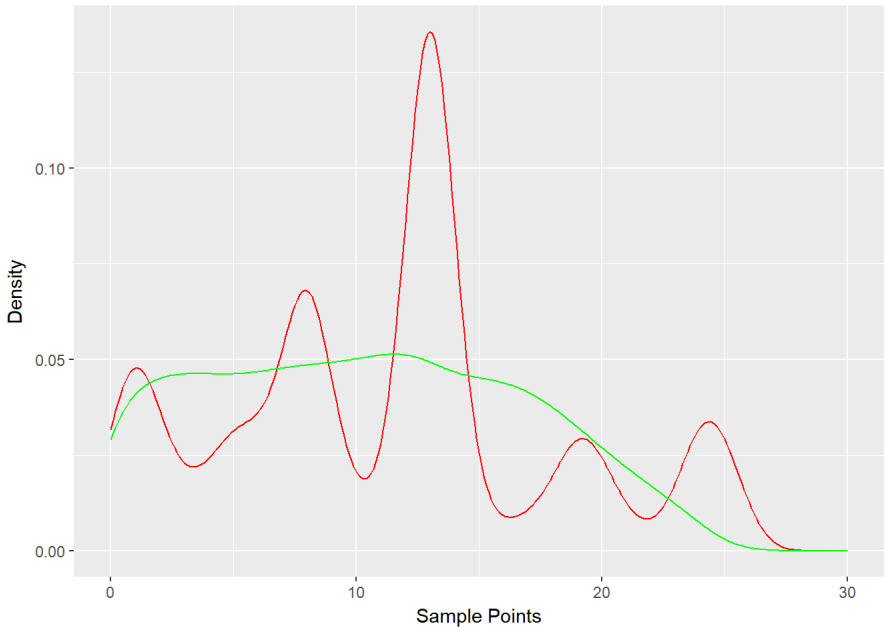
5.4 :Does Age affect termination?



Interpretation: Seeing the box plot we can say that the median age of employees whose status is active is around 40 whereas the ones which have been terminated have an average age of 60. This shows that higher ages leads to termination more often.Therfore, we can say that Age is an important factor in termination.

5.5 Does Length of Service affect termination?

```
gauss <- function(x) {
  y <-1/sqrt(2*pi) * exp(-(x^2)/2)
  return(y)
}
kde_func<-function(x,status){
  col<-c()
  stat <-dplyr::filter(new_terminations,STATUS==status)
  obv_len <- dplyr::select(stat,length_of_service)
  for (j in  obv_len){
      diff<-x-j
      density_estimates <-gauss(diff)
      col<-c(col,density_estimates)
      diff<-0
      density_estimates<-0
  }
  summ<-sum(col)
  est<-summ/nrow(obv_len)
  return(est)
}
  x<-sample(seq(0,30,0.1))
  active_col<-c()
  term_col<-c()
  for(i in x){
    active_col<-c(active_col,kde_func(i,"ACTIVE"))
    term_col<-c(term_col,kde_func(i,"TERMINATED"))
  }
  active_df<-data.frame(active_col)
  term_df<-data.frame(term_col)
  df<-cbind(active_df,term_df)
  ggplot(df, aes(x)) +
  geom_line(aes(y=term_col), colour="red") +
  geom_line(aes(y=active_col), colour="green")+
  xlab("Sample Points")+ylab("Density")
```

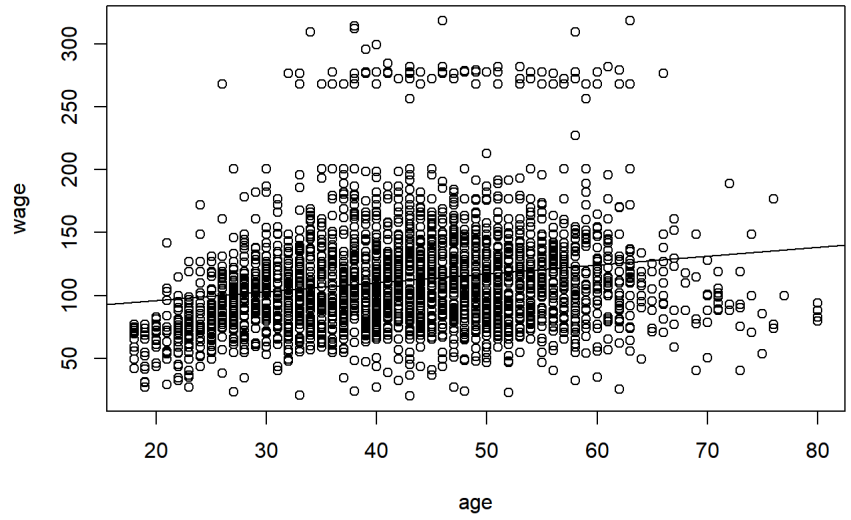


Interpretation: It can be deduced that the red line(terminated employees) has more bumps which menas that length of service sees alot of variation among terminated employees

Question 6: Applied Linear Regression

6.1: Regressing wage on age:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	81.7047354	2.8462422	28.70618	0
age	0.7072759	0.0647511	10.92299	0



6.2:Regressing Wage on age,jobclass,interraction of job class and age

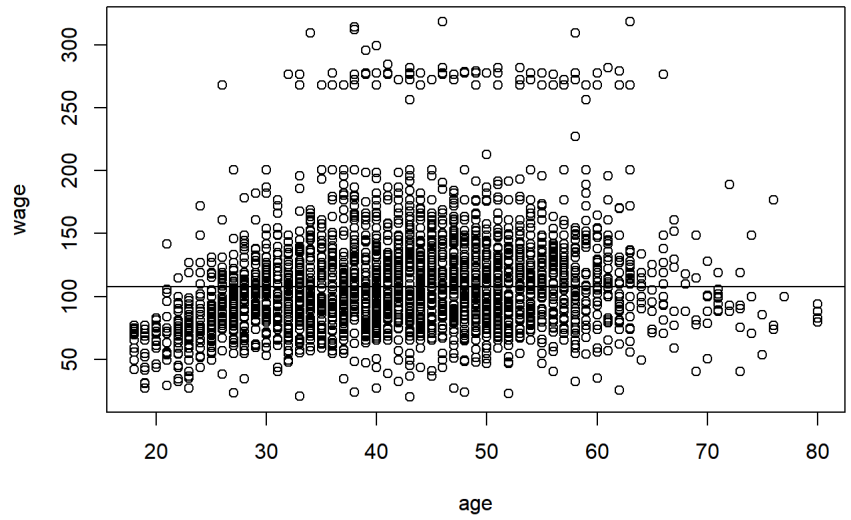
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.5283101	3.7613315	19.548479	0.0000000
age	0.7196626	0.0874392	8.230438	0.0000000
jobclass2. Information	22.7308551	5.6314068	4.036443	0.0000556
age:jobclass2. Information	-0.1601733	0.1278514	-1.252808	0.2103733

Interpretation:

1. We see that for every 1 year increase in age, and keeping other parameters constant wage increases by a factor of 0.07.
2. We also say that, for every increase of 1 level in job class increases wage by a factor of 22.7
3. We also can comment that if there is an increase of 1 $ and an increase of 1 level of job class the wage actually decreases by factor of 0.16

6.3:Regressing wage on 4th degree polynomial of age:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	108.0365847	1.0541486	102.48705	0e+00
I(age^4)	0.0000008	0.0000002	5.01142	6e-07



Interpretation: It shows that, in a fourth degree non-linear regression the wage almost remains same(horizontal fitted line) across age groups(20-60 years).

6.4:Regressing wage with all parameters

```
## The following objects are masked from Wage (pos = 3):
##
##     age, education, health, health_ins, jobclass, logwage, maritl,
##     race, region, wage, year
```

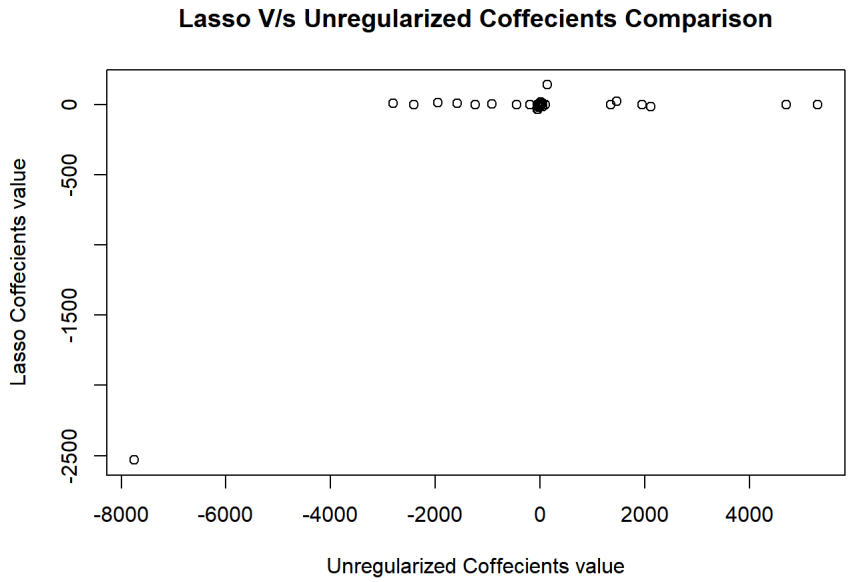
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7753.9878320	3.527567e+03	-2.1981120	0.0280205
year	3.9064463	1.759254e+00	2.2205127	0.0264616
age	105.1818000	6.209706e+01	1.6938289	0.0904060
maritl2. Married	-1954.8294255	1.718873e+03	-1.1372741	0.2555184
maritl3. Widowed	-2403.9842513	2.239426e+04	-0.1073482	0.9145202
maritl4. Divorced	-1577.1764114	2.967512e+03	-0.5314811	0.5951264
maritl5. Separated	-1239.2926357	4.965046e+03	-0.2496034	0.8029118
race2. Black	-446.9239880	2.263837e+03	-0.1974188	0.8435138
race3. Asian	5305.8656618	2.507827e+03	2.1157225	0.0344541
race4. Other	4704.0871022	6.004561e+03	0.7834190	0.4334456
education2. HS Grad	-194.1489520	2.340968e+03	-0.0829353	0.9339087
education3. Some College	1357.6269422	2.506166e+03	0.5417148	0.5880569
education4. College Grad	1470.3987000	2.517487e+03	0.5840740	0.5592163
education5. Advanced Degree	-2806.4464414	2.766035e+03	-1.0146099	0.3103771
jobclass2. Information	-917.9961304	1.327927e+03	-0.6913000	0.4894328
health2. >=Very Good	1950.9107872	1.440992e+03	1.3538665	0.1758853
health_ins2. No	2117.1199039	1.398057e+03	1.5143300	0.1300519
year:age	-0.0524606	3.095740e-02	-1.6946057	0.0902584
year:maritl2. Married	0.9867612	8.570900e-01	1.1512924	0.2497076
year:maritl3. Widowed	1.1776423	1.113896e+01	0.1057228	0.9158096
year:maritl4. Divorced	0.8014230	1.479802e+00	0.5415743	0.5881537

	Estimate	Std. Error	t value	Pr(>|t|)
year:maritl5. Separated	0.6193224	2.474723e+00	0.2502593	0.8024047
year:race2. Black	0.2252702	1.129206e+00	0.1994944	0.8418902
year:race3. Asian	-2.6474321	1.251165e+00	-2.1159742	0.0344326
year:race4. Other	-2.3425831	2.992688e+00	-0.7827689	0.4338272
year:education2. HS Grad	0.0978807	1.167770e+00	0.0838185	0.9332066
year:education3. Some College	-0.6810055	1.250115e+00	-0.5447543	0.5859647
year:education4. College Grad	-0.7239778	1.255845e+00	-0.5764866	0.5643314
year:education5. Advanced Degree	1.4014164	1.379568e+00	1.0158372	0.3097923
year:jobclass2. Information	0.4598472	6.621812e-01	0.6944432	0.4874604
year:health2. >=Very Good	-0.9764438	7.185605e-01	-1.3588887	0.1742883
year:health_ins2. No	-1.0655096	6.971138e-01	-1.5284586	0.1265086
age:maritl2. Married	-0.2208946	1.608386e-01	-1.3733932	0.1697371
age:maritl3. Widowed	0.1690018	2.074471e+00	0.0814674	0.9350759
age:maritl4. Divorced	-0.3672340	2.967717e-01	-1.2374291	0.2160287
age:maritl5. Separated	0.0419255	5.445532e-01	0.0769907	0.9386363
age:race2. Black	0.0039151	2.064997e-01	0.0189593	0.9848749
age:race3. Asian	-0.0660929	2.805349e-01	-0.2355959	0.8137630
age:race4. Other	0.2141821	5.880956e-01	0.3641961	0.7157383
age:education2. HS Grad	0.0785983	2.333801e-01	0.3367825	0.7363054
age:education3. Some College	0.6547585	2.519756e-01	2.5985001	0.0094110
age:education4. College Grad	0.1670510	2.570979e-01	0.6497563	0.5159015
age:education5. Advanced Degree	0.5396877	2.857042e-01	1.8889733	0.0589958
age:jobclass2. Information	0.0123985	1.335954e-01	0.0928060	0.9260641
age:health2. >=Very Good	0.3139447	1.376458e-01	2.2808149	0.0226322
age:health_ins2. No	0.1616726	1.387698e-01	1.1650412	0.2440989
maritl2. Married:race2. Black	-5.8332749	5.411073e+00	-1.0780256	0.2811126
maritl3. Widowed:race2. Black	-43.6363626	1.156695e+02	-0.3772505	0.7060152
maritl4. Divorced:race2. Black	-13.3155910	9.549267e+00	-1.3944097	0.1633015
maritl5. Separated:race2. Black	2.6930474	1.457367e+01	0.1847886	0.8534079
maritl2. Married:race3. Asian	-7.4912662	7.569605e+00	-0.9896509	0.3224279
maritl3. Widowed:race3. Asian	-16.9650373	4.597916e+01	-0.3689723	0.7121755
maritl4. Divorced:race3. Asian	-18.9820416	3.658917e+01	-0.5187885	0.6039481
maritl5. Separated:race3. Asian	-12.3130625	4.204224e+01	-0.2928736	0.7696399
maritl2. Married:race4. Other	-11.2507790	1.455441e+01	-0.7730149	0.4395770
maritl3. Widowed:race4. Other	-12.4774933	5.737621e+01	-0.2174680	0.8278590
maritl4. Divorced:race4. Other	-27.7983651	2.613524e+01	-1.0636352	0.2875832
maritl2. Married:education2. HS Grad	4.1637946	6.690017e+00	0.6223892	0.5337353
maritl3. Widowed:education2. HS Grad	52.9331618	1.011294e+02	0.5234202	0.6007222
maritl4. Divorced:education2. HS Grad	-3.9054177	1.133798e+01	-0.3444546	0.7305296
maritl5. Separated:education2. HS Grad	16.8895748	1.456941e+01	1.1592487	0.2464510
maritl2. Married:education3. Some College	1.1067751	7.081409e+00	0.1562931	0.8758130
maritl3. Widowed:education3. Some College	27.2549641	6.165800e+01	0.4420345	0.6584974

	Estimate	Std. Error	t value	Pr(>|t|)
maritl4. Divorced:education3. Some College	-8.6959410	1.182169e+01	-0.7355922	0.4620388
maritl5. Separated:education3. Some College	6.1939769	1.752169e+01	0.3535033	0.7237370
maritl2. Married:education4. College Grad	8.2813598	7.305930e+00	1.1335120	0.2570937
maritl3. Widowed:education4. College Grad	36.4864088	5.951737e+01	0.6130380	0.5398997
maritl4. Divorced:education4. College Grad	1.2005614	1.226842e+01	0.0978578	0.9220520
maritl5. Separated:education4. College Grad	17.6256931	2.048610e+01	0.8603733	0.3896549
maritl2. Married:education5. Advanced Degree	19.2836005	8.217818e+00	2.3465598	0.0190151
maritl3. Widowed:education5. Advanced Degree	52.8796401	1.426689e+02	0.3706459	0.7109285
maritl4. Divorced:education5. Advanced Degree	-14.4288890	1.399706e+01	-1.0308516	0.3026970
maritl5. Separated:education5. Advanced Degree	145.6182648	3.955767e+01	3.6811637	0.0002364
maritl2. Married:jobclass2. Information	-3.7863952	3.729592e+00	-1.0152303	0.3100814
maritl3. Widowed:jobclass2. Information	5.7725928	6.909314e+01	0.0835480	0.9334216
maritl4. Divorced:jobclass2. Information	2.4057943	6.183919e+00	0.3890404	0.6972750
maritl5. Separated:jobclass2. Information	-14.2474510	1.130947e+01	-1.2597810	0.2078505
maritl2. Married:health2. >=Very Good	-2.2905959	3.940621e+00	-0.5812780	0.5610986
maritl3. Widowed:health2. >=Very Good	-6.8142077	6.034854e+01	-0.1129142	0.9101064
maritl4. Divorced:health2. >=Very Good	-4.3293192	6.287076e+00	-0.6886062	0.4911266
maritl5. Separated:health2. >=Very Good	-6.7798097	1.213327e+01	-0.5587786	0.5763563
maritl2. Married:health_ins2. No	-0.8056760	3.909368e+00	-0.2060885	0.8367363
maritl3. Widowed:health_ins2. No	20.3216325	3.655120e+01	0.5559771	0.5782697
maritl4. Divorced:health_ins2. No	-11.1185545	6.701998e+00	-1.6589911	0.0972265
maritl5. Separated:health_ins2. No	5.9864491	1.207270e+01	0.4958667	0.6200263
race2. Black:education2. HS Grad	-0.4508568	8.035991e+00	-0.0561047	0.9552623
race3. Asian:education2. HS Grad	-1.0929820	1.135866e+01	-0.0962245	0.9233489
race4. Other:education2. HS Grad	-6.5020599	1.602302e+01	-0.4057950	0.6849234
race2. Black:education3. Some College	-5.5421169	8.259167e+00	-0.6710262	0.5022577
race3. Asian:education3. Some College	-3.7688949	1.297462e+01	-0.2904822	0.7714683
race4. Other:education3. Some College	6.2026788	1.854900e+01	0.3343942	0.7381064
race2. Black:education4. College Grad	-9.4758399	9.235232e+00	-1.0260532	0.3049527
race3. Asian:education4. College Grad	11.6120040	1.112910e+01	1.0433915	0.2968546
race4. Other:education4. College Grad	-43.7572560	2.605196e+01	-1.6796149	0.0931408
race2. Black:education5. Advanced Degree	-11.1420166	1.059060e+01	-1.0520664	0.2928574
race3. Asian:education5. Advanced Degree	1.8298877	1.128578e+01	0.1621410	0.8712062
race4. Other:education5. Advanced Degree	-40.2090466	3.186235e+01	-1.2619612	0.2070650
race2. Black:jobclass2. Information	4.2368742	4.602318e+00	0.9205958	0.3573386
race3. Asian:jobclass2. Information	4.5270782	5.791027e+00	0.7817402	0.4344316
race4. Other:jobclass2. Information	8.8058937	1.482928e+01	0.5938179	0.5526806
race2. Black:health2. >=Very Good	-5.3035394	5.022597e+00	-1.0559357	0.2910862
race3. Asian:health2. >=Very Good	6.5153387	6.241519e+00	1.0438707	0.2966329
race4. Other:health2. >=Very Good	-9.0302538	1.258358e+01	-0.7176221	0.4730486
race2. Black:health_ins2. No	1.3554749	4.746258e+00	0.2855881	0.7752142
race3. Asian:health_ins2. No	-0.9591856	6.039116e+00	-0.1588288	0.8738149

	Estimate	Std. Error	t value	Pr(>|t|)
race4. Other:health_ins2. No	-5.9886404	1.386235e+01	-0.4320076	0.6657682
education2. HS Grad:jobclass2. Information	-3.1789041	5.481500e+00	-0.5799332	0.5620050
education3. Some College:jobclass2. Information	-1.2836585	5.677187e+00	-0.2261082	0.8211333
education4. College Grad:jobclass2. Information	1.9999099	5.690700e+00	0.3514348	0.7252879
education5. Advanced Degree:jobclass2. Information	13.7403956	6.419178e+00	2.1405226	0.0323963
education2. HS Grad:health2. >=Very Good	2.0572060	5.014563e+00	0.4102463	0.6816558
education3. Some College:health2. >=Very Good	1.0159992	5.402177e+00	0.1880722	0.8508333
education4. College Grad:health2. >=Very Good	3.6865253	5.664338e+00	0.6508307	0.5152077
education5. Advanced Degree:health2. >=Very Good	6.5424649	6.632123e+00	0.9864813	0.3239799
education2. HS Grad:health_ins2. No	-0.5151510	5.053166e+00	-0.1019462	0.9188065
education3. Some College:health_ins2. No	2.8846354	5.454109e+00	0.5288922	0.5969210
education4. College Grad:health_ins2. No	-11.0245106	5.593568e+00	-1.9709264	0.0488279
education5. Advanced Degree:health_ins2. No	-9.7253296	6.495468e+00	-1.4972484	0.1344383
jobclass2. Information:health2. >=Very Good	-0.0115168	3.047441e+00	-0.0037792	0.9969849
jobclass2. Information:health_ins2. No	-1.0305227	3.002720e+00	-0.3431964	0.7314757
health2. >=Very Good:health_ins2. No	0.3475195	3.128001e+00	0.1110995	0.9115452

6.5 Lasso Regression and Comparision between coeffecients



Interpretation: We can observe from the sctter plot that many lasso and unregularized coffecient values match at lower values around 0 (that is why it has a darker black dot at 0).Moreover, we find that Lasso coffecients values are generally surrounded at or around 0, because now the coffecient value get regularized and useless variebles are penalized to reduction in coffeceint values.

(Discussed with Karuna Saproo)