Intro to Data Science Project Report
Pay Equity Analysis

Introduction/ Motivation/ Problem Definition

The gender wage gap refers to the difference in earnings between women and men. According to studies, in 2020, women earned 84% of what men earned. There are considerable differences in pay between two individuals, who have similar - educational qualifications, relevant work experience, performance, and seniority and the only difference being the gender.

Prior Work

Pay equity has become a major problem in the recent times. Gender wage gap is commonly calculated by dividing women's wages by men's wages, and this ratio is often expressed as a percent, or in dollar terms. This tells us how much a woman is paid for each dollar paid to a man. This gender pay ratio is often measured for year-round, full-time workers.

For this project, we are trying to solve the gender wage problem for a given company by using data mining methods to visualize and evaluate the wage gap between men and women who have similar qualifications. We will also try to recommend pay increase for such individuals so that the overall pay gap index for the company is reduced and all employees are compensated fairly. As we are aiming at solving the gender gap problem for a given company, we will be needing a company's HR data to run our data mining techniques. The data should comprise typical HRIS data such as Employee ID, Department, Educational Qualifications, Age, Seniority, Performance rating, Experience in Field, Region of Work and Annual Salary. For this project, we will be using a dataset named 'HR Analytics: Employee Promotion Data' from Kaggle.

Model/ Algorithm/ Method

In this project, we are using linear models to visualize and analyze gender pay disparity. To begin with, we take a company's HR data which contains employees' demographic information. This data is cleaned and normalized. We then take the employee salary value and use employees' categorical data to draw regression lines between the two genders. This helps us visualize the amount of pay disparity in the organization. To come up with a model that suggests amount of pay required to individual employees to reduce organizational pay disparity, we first calculate the percentiles of the employees' categorical data and calculate mean of all the percentiles. This gives a rank to all employees. Before we begin to analyze pay disparity among employees, we need to keep in mind that only employees in the same department in the same location must be compared. This is because it wouldn't be right if a janitor is compared to an executive officer. Also, employees in different locations cannot be directly compared. This is because employees in different locations are paid differently. With this in mind, we calculate the ratios of averages of employees' categorical data and if the ratio is in between 0.99 and 1. We then calculate the ratio of employees' salaries. If this value is greater than or equal to 1, it means that as average score ratio is very close to the male and female employee being compared, they have very similar demographic variables, and hence,

should be paid equally. When annual salary ratio is greater than or equal to 1, it means that there is a pay gap. When all 4 of these conditions satisfy for a pair of employees, it means that the male employee and the female employee have similar demographic variables, but have gender pay gap. The user is presented with the data of both employees in question. This helps the user make an informed decision on pay increase for the female employee. As the user decreases the pay disparity by giving salary hikes to female employees, the regression lines drawn between the salary and the ranks for male and female employees get closer and closer. Finally, when the lines coincide, which is highly unlikely in real life, we may say that pay equity is achieved in the organization.

Results and Findings

After the models discussed above have been applied, we get the following findings:

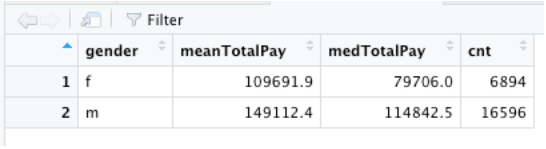
- After all the data massage has been done, the structure of the imported data is:

```
> str(data)
'data.frame':   23490 obs. of  28 variables:
 $ employee_id         : int  62237 27130 61236 24731 76338 3976 68390 5361 72735 5362 ...
 $ department          : chr  "Sales & Marketing" "Sales & Marketing" "Sales & Marketing" "Sales & Marketing" ...
 $ region              : int  9 9 9 9 9 9 9 9 9 9 ...
 $ education           : chr  "Master's & above" "Master's & above" "Master's & above" "Master's & above" ...
 $ gender              : chr  "m" "f" "f" "f" ...
 $ recruitment_channel : chr  "other" "other" "sourcing" "other" ...
 $ no_of_trainings     : int  1 2 1 1 1 2 1 1 1 1 ...
 $ age                 : int  27 32 28 28 28 33 37 36 37 32 ...
 $ previous_year_rating: int  5 1 1 1 4 1 2 3 5 3 ...
 $ length_of_service   : int  2 4 4 1 4 6 9 3 3 7 ...
 $ awards_won          : int  0 0 1 0 0 0 0 0 0 0 ...
 $ avg_training_score  : int  44 44 45 46 46 46 46 47 47 47 ...
 $ ann_salary          : int  58508 88248 41731 62047 90871 112469 56183 142828 86350 88999 ...
 $ age_bin             : num  2 2 2 2 2 2 3 3 3 2 ...
 $ dept_bin            : num  3 3 3 3 3 3 3 3 3 3 ...
 $ reg_bin             : num  6 6 6 6 6 6 6 6 6 6 ...
 $ edu_bin             : num  3 3 3 3 3 3 3 3 3 3 ...
 $ dept_perc           : num  0.487 0.487 0.487 0.487 0.487 ...
 $ reg_perc            : num  0.647 0.647 0.647 0.647 0.647 ...
 $ edu_perc            : num  0.846 0.846 0.846 0.846 0.846 ...
 $ age_perc            : num  0.304 0.304 0.304 0.304 0.304 ...
 $ perf_perc           : num  0.884 0.0653 0.0653 0.0653 0.6695 ...
 $ serv_perc           : num  0.1459 0.3981 0.3981 0.0424 0.3981 ...
 $ award_perc          : num  0.489 0.489 0.989 0.489 0.489 ...
 $ trscr_perc          : num  0.00766 0.00766 0.01784 0.03536 0.03536 ...
 $ log_total           : num  11 11.4 10.6 11 11.4 ...
 $ male                : num  1 0 0 0 1 1 1 1 1 1 ...
 $ female              : num  0 1 1 1 0 0 0 0 0 0 ...
```

- Statistical Summary of the data

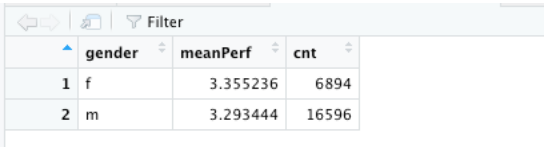
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
employee_id	23,490	39,041.40	22,640.81	3	19,370.2	58,690	78,295
region	23,490	14.29	10.15	1	4	22	34
no_of_trainings	23,490	1.25	0.60	1	1	1	9
age	23,490	34.78	7.68	20	29	39	60
previous_year_rating	23,490	3.31	1.28	1	3	4	5
length_of_service	23,490	5.81	4.21	1	3	7	34
awards_won	23,490	0.02	0.15	0	0	0	1
avg_training_score	23,490	63.26	13.41	39	51	76	99
ann_salary	23,490	137,543.00	101,569.10	40,000	73,653.5	142,487.5	549,875
age_bin	23,490	2.53	0.81	1	2	3	5
dept_bin	23,490	3.70	2.26	1	2	5	9
reg_bin	23,490	4.65	2.03	1	3	7	7
edu_bin	23,490	2.29	0.49	1	2	3	3
dept_perc	23,490	0.50	0.28	0.06	0.23	0.76	0.99
reg_perc	23,490	0.50	0.28	0.03	0.28	0.86	0.86
edu_perc	23,490	0.50	0.23	0.01	0.35	0.85	0.85
age_perc	23,490	0.50	0.26	0.02	0.30	0.73	0.99
perf_perc	23,490	0.50	0.28	0.07	0.40	0.67	0.88
serv_perc	23,490	0.50	0.29	0.04	0.27	0.71	1.00
award_perc	23,490	0.50	0.07	0.49	0.49	0.49	0.99
trscr_perc	23,490	0.50	0.29	0.0000	0.25	0.75	1.00
log_total	23,490	11.63	0.59	10.60	11.21	11.87	13.22
male	23,490	0.71	0.46	0	0	1	1
female	23,490	0.29	0.46	0	0	1	1

- Displaying Total Pay summary grouped by gender


Filter

	gender	meanTotalPay	medTotalPay	cnt
1	f	109691.9	79706.0	6894
2	m	149112.4	114842.5	16596

- Performance evaluations summary stats grouped by gender


Filter

	gender	meanPerf	cnt
1	f	3.355236	6894
2	m	3.293444	16596

- Distribution of employees by department, grouped by gender

Filter

	dept	gender	meanTotalPay	cnt
1	Technology	f	248262.55	1171
2	Technology	m	349513.24	1840
3	Sales & Marketing	f	69777.05	1373
4	Sales & Marketing	m	100398.98	5942
5	R&D	f	302236.12	25
6	R&D	m	400766.13	415
7	Procurement	f	69573.47	1359
8	Procurement	m	100039.50	1661
9	Operations	f	69660.22	1972
10	Operations	m	101397.12	2792
11	Legal	f	253658.34	47
12	Legal	m	305469.42	398
13	HR	f	98988.59	453
14	HR	m	131476.81	632
15	Finance	f	214451.72	271
16	Finance	m	251195.53	820
17	Analytics	f	68791.05	223
18	Analytics	m	99598.64	2096

- Distribution of employees by region, grouped by gender

Filter

	region	gender	meanTotalPay	cnt
1	34	f	65426.82	11
2	34	m	117474.83	144
3	33	f	215912.09	23
4	33	m	226361.21	103
5	32	f	90233.60	91
6	32	m	133440.70	342
7	31	f	102580.94	219
8	31	m	139821.88	625
9	30	f	107696.85	67
10	30	m	129552.76	206
11	29	f	100827.50	70
12	29	m	134947.90	344
13	28	f	111445.73	128
14	28	m	124042.46	467
15	27	f	113086.97	205
16	27	m	164004.37	505
17	26	f	117842.87	248
18	26	m	171407.09	763
19	25	f	113595.72	68
20	25	m	131709.97	269
21	24	f	129341.82	60
22	24	m	166893.86	159
23	23	f	103581.90	125
24	23	m	161120.43	391
25	22	f	99916.99	740
26	22	m	154957.89	1999
27	21	f	98119.43	37

Running the Regression models

- Using No controls.

```
> reg1 <- lm(log_total ~ male, data = data)
> summary(reg1)

Call:
lm(formula = log_total ~ male, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90777 -0.40363 -0.09032  0.15778  1.59417

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.406878   0.006899 1653.52   <2e-16 ***
male         0.320773   0.008207   39.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5728 on 23488 degrees of freedom
Multiple R-squared:  0.06107,   Adjusted R-squared:  0.06103
F-statistic:  1528 on 1 and 23488 DF,  p-value: < 2.2e-16
```

- Using performance rating, age and highest education as controls

```
> reg2 <- lm(log_total ~ male + previous_year_rating + age_bin + education, data = data)
> summary(reg2)

Call:
lm(formula = log_total ~ male + previous_year_rating + age_bin +
    education, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22060 -0.40236 -0.08894  0.16116  1.58527

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               11.449709   0.016817 680.857   <2e-16 ***
male                       0.321135   0.008191  39.207   <2e-16 ***
previous_year_rating      -0.003485   0.002915  -1.196   0.2319
age_bin                   -0.015754   0.004899  -3.216   0.0013 **
educationBelow Secondary   0.303915   0.030299  10.030   <2e-16 ***
educationMaster's & above  0.011521   0.008443   1.365   0.1724
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5713 on 23484 degrees of freedom
Multiple R-squared:  0.0661,    Adjusted R-squared:  0.0659
F-statistic: 332.4 on 5 and 23484 DF,  p-value: < 2.2e-16
```

- Using employees' whole demographic data as control

```
Call:
lm(formula = log_total ~ male + previous_year_rating + age_bin +
    education + department + length_of_service + avg_training_score,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91217 -0.23620  0.04381  0.25935  0.70205

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                11.0366354  0.0378793 291.363   <2e-16 ***
male                        0.3446069  0.0045479  75.772   <2e-16 ***
previous_year_rating        0.0011792  0.0015763   0.748   0.4544
age_bin                     0.0029165  0.0031726   0.919   0.3580
educationBelow Secondary    0.0150936  0.0163007   0.926   0.3545
educationMaster's & above  -0.0019419  0.0045262  -0.429   0.6679
departmentFinance           0.9694035  0.0153426  63.184   <2e-16 ***
departmentHR                0.3320538  0.0186547  17.800   <2e-16 ***
departmentLegal             1.1544821  0.0191370  60.327   <2e-16 ***
departmentOperations        0.0334115  0.0131450   2.542   0.0110 *
departmentProcurement       0.0162326  0.0106477   1.525   0.1274
departmentR&D               1.4164638  0.0158499  89.367   <2e-16 ***
departmentSales & Marketing 0.0381517  0.0165719   2.302   0.0213 *
departmentTechnology        1.2596860  0.0088090 143.000   <2e-16 ***
length_of_service          -0.0004846  0.0005873  -0.825   0.4093
avg_training_score          0.0008583  0.0004322   1.986   0.0471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3039 on 23474 degrees of freedom
Multiple R-squared:  0.7358,    Adjusted R-squared:  0.7356
F-statistic:  4358 on 15 and 23474 DF,  p-value: < 2.2e-16
```

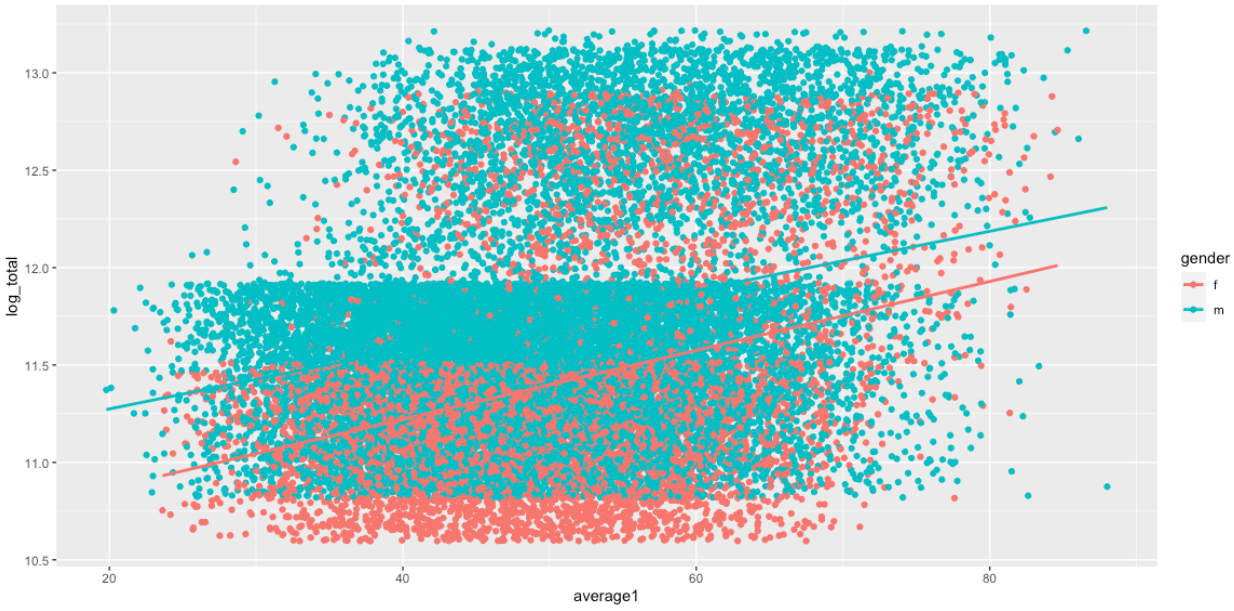
- HTML Output of the Regression results. Adjusted pay gap here is 34.5%

	Dependent variable:		
	log_total		
	(1)	(2)	(3)
male	0.321***	0.321***	0.345***
	(0.008)	(0.008)	(0.005)
previous_year_rating		-0.003	0.001
		(0.003)	(0.002)
age_bin		-0.016***	0.003
		(0.005)	(0.003)
educationBelow Secondary		0.304***	0.015
		(0.030)	(0.016)
educationMaster's	above		0.012
		(0.008)	(0.005)
departmentFinance			0.969***
			(0.015)
departmentHR			0.332***
			(0.019)
departmentLegal			1.154***
			(0.019)
departmentOperations			0.033**
			(0.013)
departmentProcurement			0.016
			(0.011)
departmentR&D			1.416***
			(0.016)
departmentSales	Marketing		
			(0.017)
departmentTechnology			1.260***
			(0.009)

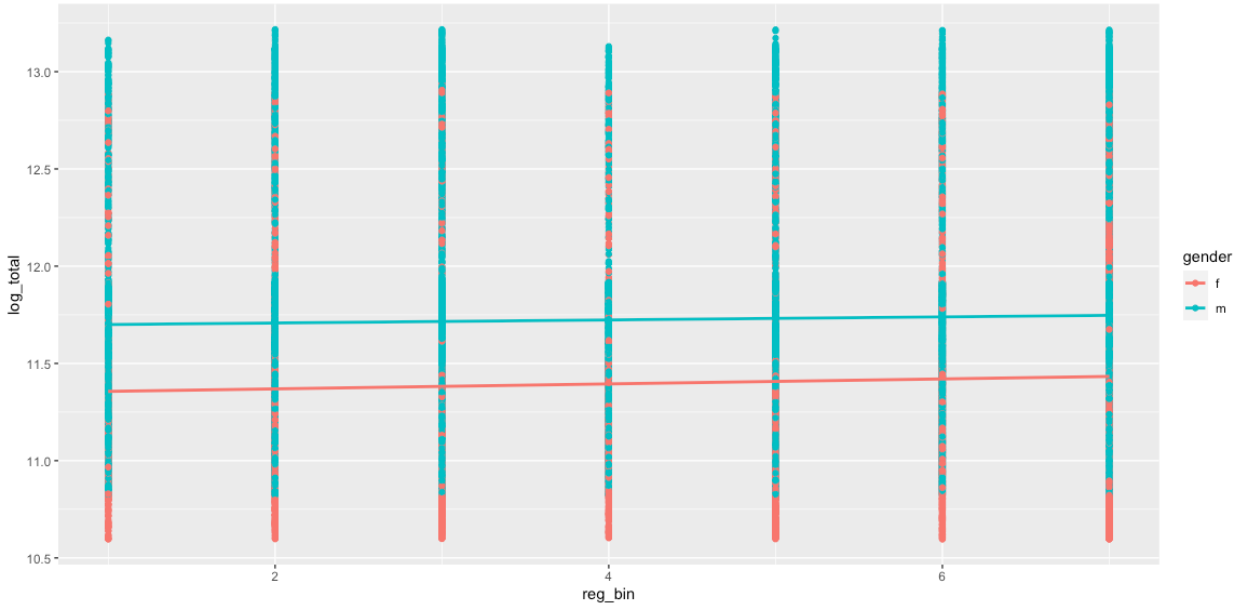
- Department wise results of the linear model.

	A	B	C	D	E	F	G
		term	estimate	std.error	statistic	p.value	
1		(Intercept)	11.20374543	0.042288459	264.936241	0	
2		male	0.306205875	0.021394542	14.3123358	2.8705E-46	
3		departmentFinance	1.023745413	0.029378404	34.8468701	0.000000000000000(	
4		departmentHR	0.297006308	0.028984514	10.2470687	1.3745E-24	
5		departmentLegal	1.131746476	0.049928583	22.6673063	1.523E-112	
6		departmentOperations	-0.056726651	0.023955808	-2.3679707	0.01789401	
7		departmentProcurement	-0.051044215	0.022887381	-2.230234	0.02574133	
8		departmentR&D	1.355348049	0.064047595	21.161576	1.7969E-98	
9		departmentSales & Marketing	-0.048287146	0.02660605	-1.8148934	0.06955305	
10		departmentTechnology	1.196571284	0.022342418	53.5560333	0	
11		previous_year_rating	0.001200737	0.001578214	0.76081978	0.44677236	
12		age_bin	0.004617869	0.003169668	1.45689354	0.14515915	
13		educationBelow Secondary	0.011407593	0.016287054	0.70040859	0.48367912	
14		educationMaster's & above	0.000924837	0.004521236	0.20455401	0.83792237	
15		length_of_service	-0.000150476	0.000586722	-0.2564689	0.79759105	
16		avg_training_score	-0.000688711	0.000431713	-1.595297	0.1106593	
17		male:departmentFinance	-0.155237282	0.030172227	-5.1450389	2.6962E-07	
18		male:departmentHR	-0.055373261	0.028411564	-1.9489691	0.05131101	
19		male:departmentLegal	-0.059770993	0.051475372	-1.1611571	0.24558984	
20		male:departmentOperations	0.046199279	0.02318377	1.99274226	0.04630124	
21		male:departmentProcurement	0.04560548	0.02410701	1.89179328	0.05853083	
22		male:departmentR&D	0.034683978	0.066071547	0.52494574	0.59962587	
23		male:departmentSales & Marketing	0.026389568	0.023242014	1.13542518	0.25620881	
24		male:departmentTechnology	0.063850634	0.024218864	2.63640093	0.00838458	

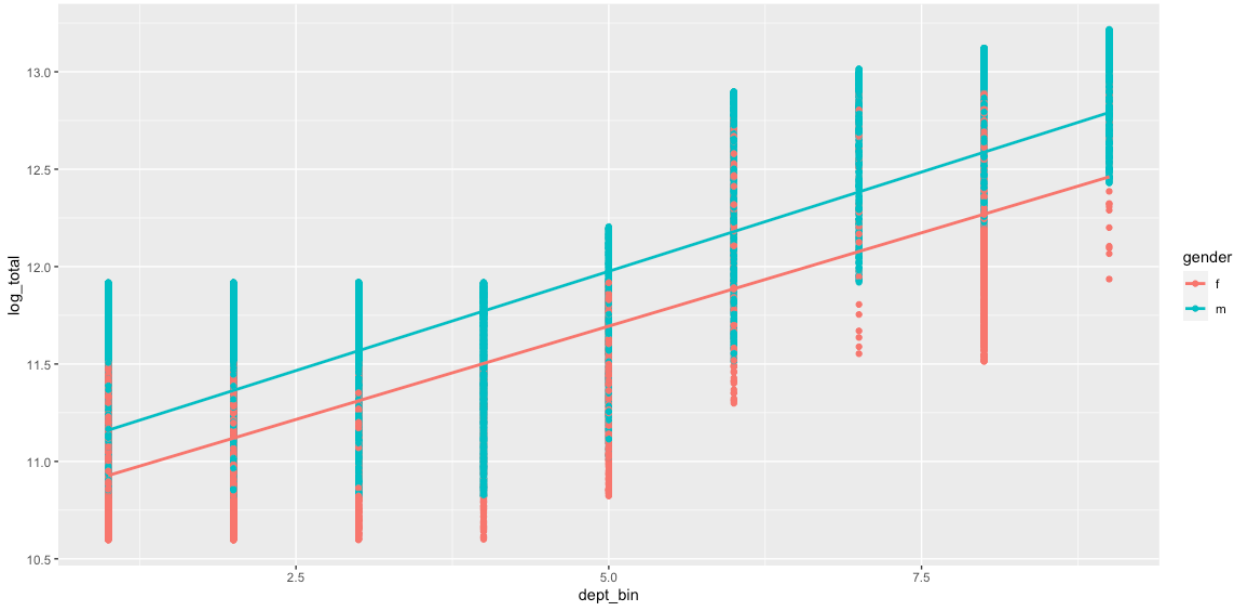
- Regression lines between log(total salary) and ranks



- Regression lines between log(total salary) and region



- Regression lines between log(total salary) and department



- This is the final output spreadsheet. This contains employee categorical and suggested increase dollar amount.

A	B	C	D	E	F	G	H	I	J	K	L
	m_emp_id	f_emp_id	m_dept	f_dept	m_region	f_region	m_average	f_average	m_ann_sal	f_ann_sal	Suggst_Amt
1	28382	27130	Sales & Marl	Sales & Marl	9	9	40.5438484	40.5467752	88753	88248	505
2	43593	70958	Sales & Marl	Sales & Marl	9	9	46.4492869	46.4492869	65539	53550	11989
3	78162	70958	Sales & Marl	Sales & Marl	9	9	46.4492869	46.4492869	146244	53550	92694
4	17233	18968	Sales & Marl	Sales & Marl	9	9	42.9584398	42.9924968	127234	53597	73637
5	18914	61236	Sales & Marl	Sales & Marl	9	9	46.8723393	46.923957	141111	41731	99380
6	6846	70958	Sales & Marl	Sales & Marl	9	9	46.0246381	46.4492869	101477	53550	47927
7	77785	43324	Analytics	Analytics	9	9	65.1551192	65.2761814	65389	64209	1180
8	6646	61236	Sales & Marl	Sales & Marl	9	9	46.8228501	46.923957	61935	41731	20204
9	67807	18968	Sales & Marl	Sales & Marl	9	9	42.6343657	42.9924968	57331	53597	3734
10	31679	61236	Sales & Marl	Sales & Marl	9	9	46.735845	46.923957	58257	41731	16526
11	47070	59733	Operations	Operations	8	8	55.4493934	55.8506279	139096	87518	51578
12	43668	47319	Operations	Operations	8	8	56.2771392	56.3944764	141501	91753	49748
13	67228	46517	Operations	Operations	8	8	51.3034802	51.5373563	125690	93301	32389
14	67228	62036	Operations	Operations	8	8	51.3034802	51.7089719	125690	74244	51446
15	73446	76477	Operations	Operations	8	8	50.2506386	50.6529374	103206	74612	28594
16	19802	22479	Operations	Operations	8	8	49.6886973	50.0119732	111990	87578	24412
17	19802	45212	Operations	Operations	8	8	49.6886973	50.1737441	111990	97205	14785
18	77651	1135	Operations	Operations	8	8	53.1800766	53.4328438	82066	75656	6410
19	77651	45458	Operations	Operations	8	8	53.1800766	53.5576309	82066	78625	3441
20	63507	34467	Procurement	Procurement	8	8	48.2468604	48.6680502	95534	94525	1009
21	41451	47319	Operations	Operations	8	8	56.3114623	56.3944764	122455	91753	30702
22	28253	8149	Procurement	Procurement	8	8	56.3963389	56.788261	85230	67544	17686
23	28253	31124	Procurement	Procurement	8	8	56.3963389	56.9106535	85230	49134	36096
24	75635	2725	Procurement	Procurement	8	8	56.8414751	57.1695402	114257	99801	14456
25	75635	31124	Procurement	Procurement	8	8	56.8414751	56.9106535	114257	49134	65123
26	75635	34894	Procurement	Procurement	8	8	56.8414751	57.0647084	114257	67592	46665
27	9050	34467	Procurement	Procurement	8	8	48.2572371	48.6680502	111535	94525	17010
28	67910	64759	Sales & Marl	Sales & Marl	7	7	45.9283206	46.1571413	60730	44045	16685
29	65914	57748	Sales & Marl	Sales & Marl	7	7	65.8777671	65.9456152	77046	41361	35685
30	50283	37091	Sales & Marl	Sales & Marl	7	7	53.8609515	54.0490634	82231	42115	40116
31	56236	33445	Sales & Marl	Sales & Marl	7	7	58.8723925	59.4396552	71537	53114	18423
32	42856	827	Sales & Marl	Sales & Marl	7	7	55.414006	55.8304066	115018	60353	54665
33	42856	35857	Sales & Marl	Sales & Marl	7	7	55.414006	55.6909855	115018	63429	51589
34	50379	10431	Sales & Marl	Sales & Marl	7	7	63.0773734	63.2763942	106133	68485	37648
35	14580	38884	Sales & Marl	Sales & Marl	7	7	42.8522776	42.9584398	104122	63735	40387
36	18182	58450	Sales & Marl	Sales & Marl	7	7	48.226639	48.2888995	103787	61322	42465