Study of Career 2030 Program's Impact on Employee Promotion Despite Treatment Corruption

Lanston Chen, Wenxi Xu, Icy Wang, Yihua Wang, Yizhou Sun

Executive Summary

In order to make data-informed decision-making and seek insights into Career 2030 program's impact on employee promotion and retention, despite the corruption of data in the treatment process that voids it as a directly usable randomized control trial (RCT), several causal inference tests that only assume observational data were conducted:

1	**One-to-one Matching**: Matches treated and control units one-to-one based on observed covariates, aiming to reduce confounding bias in observational studies
2	**Propensity Score Matching**: Matches individuals based on their propensity scores, summarizing the likelihood of receiving treatment given observed covariates, to balance treatment and control groups in observational studies
3	**IPTW**: Estimates causal effects by weighting observations based on the inverse probability of treatment, addressing confounding bias in observational studies
4	**Instrumental Variable**: Utilizes an external variable (instrument) to estimate causal effects in the presence of unobserved confounding, assuming the instrument affects treatment but not directly the outcome

While all methods yield a statistically significant result and generally pass assumption tests, further comparison of their theoretical robustness, sensitivity, and procedural compliance under our specific data will be conducted towards the end of the report – building up to both a comprehensive conclusion of the causal effect in interest and recommendation for avoiding corruption in the future.

1. Background and Exploratory Data Analysis

A randomized control trial (RCT) was designed to study the causal relation between receiving a particular professional training (Career 2030) and whether the trainee got promoted, with 2 treatment groups randomly assigned to take/ not take the training. However, two patterns in the data collected revealed that it's no longer appropriate to use the data as RCT.

For one, the 2 treatment groups were designed to be balanced at the beginning of this experiment (each account for 5% of the population), but in current data there's an unbalanced combination of 3709 not-treated and 2291 treated in the training variable. The following factors can contribute to the observed discrepancy:

- **Non-Compliance:** Some employees who were selected for training may not have complied with the program requirements, leading to a smaller proportion of employees actually receiving the training compared to the intended 5%.
- **Attrition:** Employees who left the company during the period between randomization and the availability of data may have contributed to the unequal distribution between the trained and untrained groups. If attrition rates differed between the two groups, this could further skew the distribution.
- **Managerial Intervention:** Some managers may have intervened to ensure that their direct reports were included in the training program. This intervention could have skewed the distribution of employees between the trained and untrained groups.
- **Employee Motivation:** Motivated employees who are actively seeking career development opportunities may have been more inclined to participate in the training program. This self-selection bias could result in a disproportionate number of motivated employees in the trained group.

Secondly, the randomized assignment of treatment should have controlled for all other covariates which means the distribution of other covariates should be the same for the 2 treatment groups. Yet, after checking distributions of other covariates, general imbalances in weight, distance from home, test score, participation in the flexible spending account program, and participation in 401k can be observed. (See Appendix)

Fortunately, considering the data as an observational study, causal inference techniques like control-group matching can still be conducted to test the direction and significance of causal effect – which will be the goal of this report.

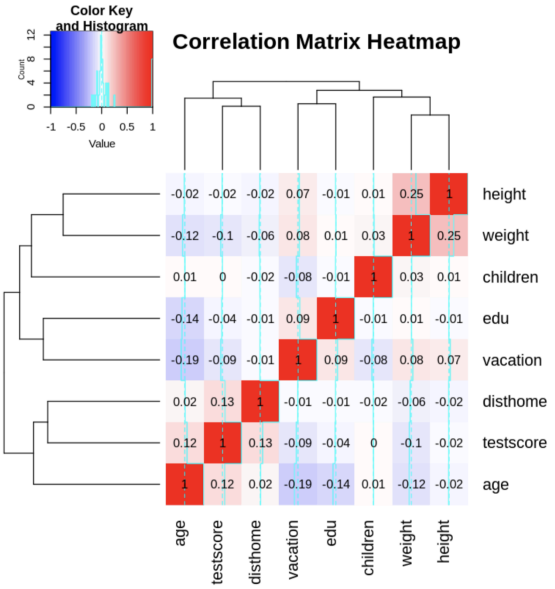
2. Data Pre-processing

The unique ID column 'empid' is dropped as it's not a meaningful confounder to be considered in matching.

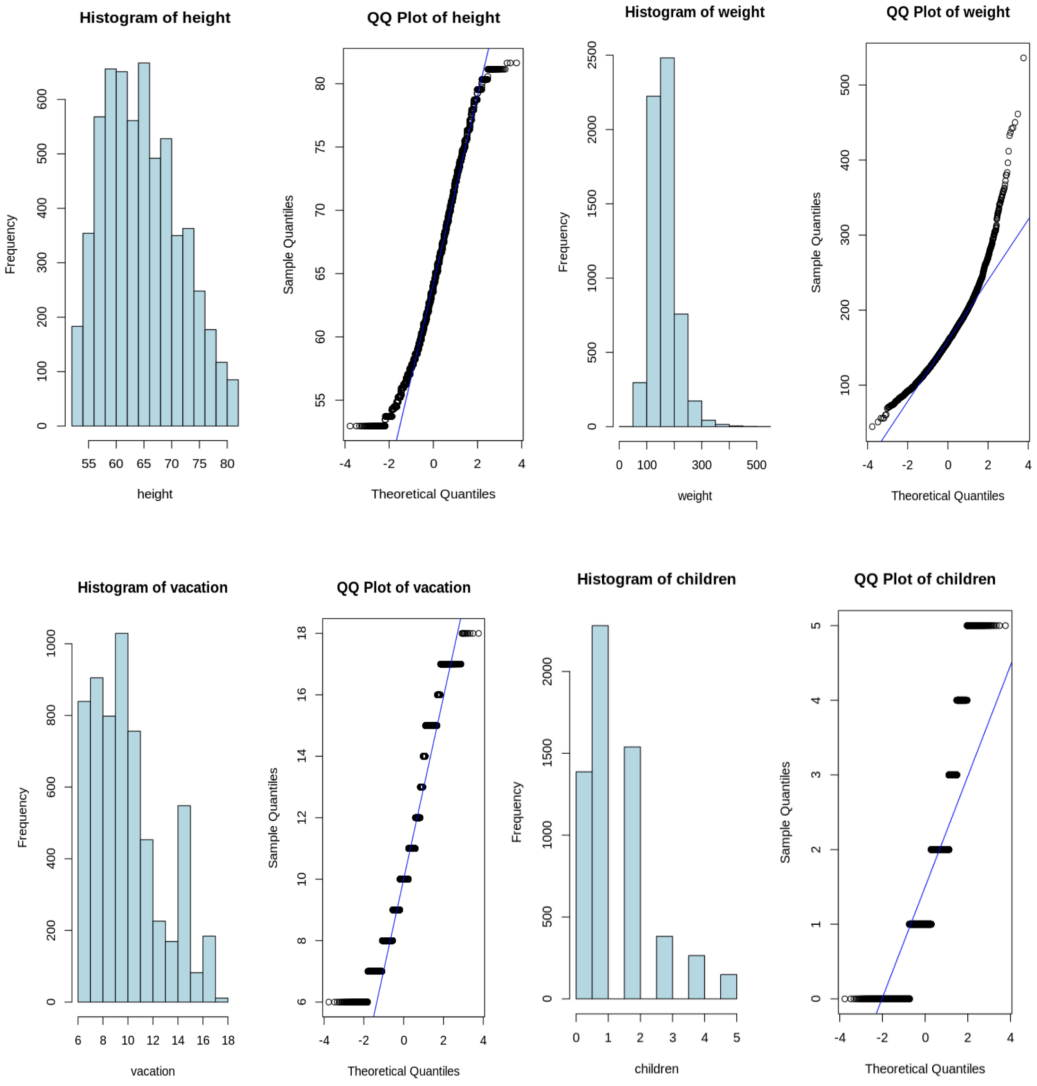
T = training
Y = promoted
Xs = all other variables as potential confounders

Transforming variables (for regression-related analysis only):

- No high correlation between variables is observed, thus passing the no-collinearity requirement
- ‘weight’, ’height’, ‘vacation’, ‘children’ are skewed and after experimenting with various combinations, the following transformation will be conducted before using regression:
    - log(weight)
    - log(height)
    - log(vacation)
    - sqrt(children)

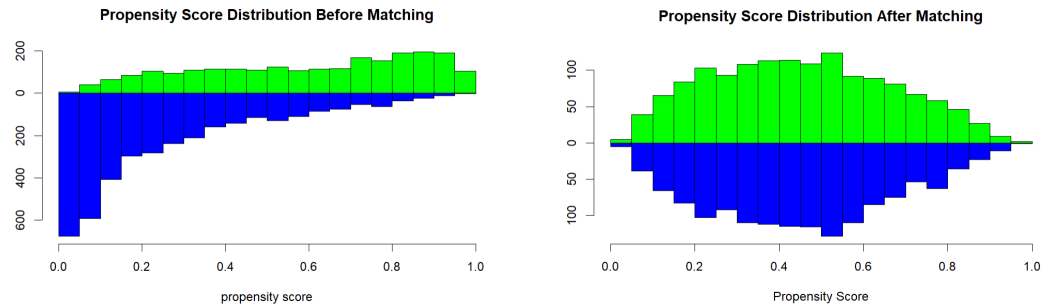


Skewness of the variables before transformation:



3. Pre-Matching RCT Data Assessment

3.1 comparing propensity score distribution:



Firstly, we analyzed the distribution of propensity scores in the original dataset. The distribution plots before matching revealed a significant asymmetry between the test and train groups' propensity scores. This asymmetry suggests that the original data does not comply with the assumptions of a randomized controlled trial (RCT).

Consequently, as a preliminary step, we applied Propensity Score Matching (PSM) using all features to the original data, resulting in a post-matching data distribution that is much more symmetrical. However, the distribution is overly broad, indicating that the matching was not optimal. Further discussion and improvement of the matching process will be addressed in the subsequent matching section.

3.2 SMD details

	Stratified by training		
	No	Yes	SMD
n	3709	2291	
manager = Yes (%)	485 (13.1)	379 (16.5)	0.098
raise = Yes (%)	1540 (41.5)	664 (29.0)	0.265
salary (%)			0.134
> $80k	269 (7.3)	204 (8.9)	
$20-$40k	743 (20.0)	471 (20.6)	
$40-$80k	531 (14.3)	412 (18.0)	
Under $20k	2166 (58.4)	1204 (52.6)	
children (mean (SD))	1.35 (1.16)	1.44 (1.21)	0.078
mstatus (%)			0.133
divorced	910 (24.5)	467 (20.4)	
married	724 (19.5)	554 (24.2)	
single	2075 (55.9)	1270 (55.4)	
age (mean (SD))	43.26 (11.96)	42.56 (10.80)	0.061
sex = Male (%)	1998 (53.9)	1340 (58.5)	0.093
edu (mean (SD))	11.51 (3.15)	11.80 (3.17)	0.093
vacation (mean (SD))	10.31 (2.85)	10.50 (2.66)	0.069
weight (mean (SD))	158.41 (45.10)	169.58 (44.10)	0.250
height (mean (SD))	64.48 (6.65)	64.91 (6.85)	0.063
hrfriend = Yes (%)	2001 (53.9)	1215 (53.0)	0.018
cxofriend = Yes (%)	1985 (53.5)	1318 (57.5)	0.081
insurance (%)			0.231
Covered	1236 (33.3)	940 (41.0)	
Covered & Medicaid	640 (17.3)	393 (17.2)	
Covered & Medicare	27 (0.7)	10 (0.4)	
Medicaid	1517 (40.9)	789 (34.4)	
Medicare	85 (2.3)	16 (0.7)	
Medicare & Medicaid	14 (0.4)	1 (0.0)	
Other	190 (5.1)	142 (6.2)	
flexspend = Yes (%)	1057 (28.5)	957 (41.8)	0.281
retcont = Yes (%)	594 (16.0)	131 (5.7)	0.335
race (%)			0.041

Secondly, for the original dataset, the standardized mean differences (SMD) for each variable were displayed using the "TableOne" library in R. Some variables, such as race, weight, insurance, discretionary spending, and restructuring, exhibited SMDs greater than 0.2. This indicates that the initial study groups were not perfectly matched. In other words, the initial A/B test did not conform to the typical randomized controlled trial.

By comparing the propensity score distributions and analyzing the Standardized Mean Differences (SMD) using Table One, we can corroborate that the original RCT does not meet the standard RCT assumptions, necessitating further matching enhancements.

4. One to One Matching

4.1 Processing

We aim to reduce confounding bias in observational studies by employing 1:1 matching. We start with the preparation of covariates, transforming categorical variables into dummy variables to facilitate numerical matching. This step ensures that diverse characteristics across individuals work effectively.

We evaluate the quality of the matches by comparing baseline characteristics between the treated and control groups. The use of Standardized Mean Differences (SMD) as a metric allows for the assessment of balance across all observed covariates.

Then, we tried tuning the Caliper as the sensitivity analysis in order to find the proper matching in this 1:1 phase. We put into different values to check whether the matching would generate a small SMD for each covariate, the causal assumption is violated, or we have enough match pairs. Below are matching tables with caliper 1 and 3:

	Stratified by training		
	No	Yes	SMD
n	82	82	
manager = Yes (%)	4 (4.9)	4 (4.9)	<0.001
raise = Yes (%)	22 (26.8)	22 (26.8)	<0.001
salary (%)			<0.001
Under $20k	70 (85.4)	70 (85.4)	
$20-$40k	5 (6.1)	5 (6.1)	
$40-$80k	6 (7.3)	6 (7.3)	
> $80k	1 (1.2)	1 (1.2)	
children (mean (SD))	1.28 (1.18)	1.33 (1.13)	0.042
mstatus (%)			<0.001
divorced	9 (11.0)	9 (11.0)	
married	11 (13.4)	11 (13.4)	
single	62 (75.6)	62 (75.6)	
age (mean (SD))	46.41 (9.53)	46.20 (9.20)	0.023
sex = Male (%)	54 (65.9)	54 (65.9)	<0.001
edu (mean (SD))	10.79 (2.02)	10.70 (2.21)	0.046
vacation (mean (SD))	9.89 (1.95)	9.83 (1.96)	0.031
weight (mean (SD))	157.85 (31.79)	157.48 (33.11)	0.011
height (mean (SD))	64.40 (5.75)	64.08 (5.91)	0.055
hrfriend = Yes (%)	45 (54.9)	45 (54.9)	<0.001
cxofriend = Yes (%)	44 (53.7)	44 (53.7)	<0.001
insurance (%)			<0.001
Covered	13 (15.9)	13 (15.9)	
Covered & Medicaid	14 (17.1)	14 (17.1)	
Covered & Medicare	0 (0.0)	0 (0.0)	
Medicaid	55 (67.1)	55 (67.1)	
Medicare	0 (0.0)	0 (0.0)	
Medicare & Medicaid	0 (0.0)	0 (0.0)	
Other	0 (0.0)	0 (0.0)	
flexspend = Yes (%)	37 (45.1)	37 (45.1)	<0.001
retcont = Yes (%)	2 (2.4)	2 (2.4)	<0.001
race (%)			<0.001
black	8 (9.8)	8 (9.8)	
other	0 (0.0)	0 (0.0)	
white	74 (90.2)	74 (90.2)	
disthome (mean (SD))	21.72 (7.55)	20.26 (8.01)	0.188
testscore (mean (SD))	61.21 (11.17)	60.79 (9.45)	0.040

Caliper = 1

	Stratified by training		
	No	Yes	SMD
n	2291	2291	
manager = Yes (%)	361 (15.8)	379 (16.5)	0.021
raise = Yes (%)	864 (37.7)	664 (29.0)	0.186
salary (%)			0.071
Under $20k	1278 (55.8)	1204 (52.6)	
$20-$40k	443 (19.3)	471 (20.6)	
$40-$80k	366 (16.0)	412 (18.0)	
> $80k	204 (8.9)	204 (8.9)	
children (mean (SD))	1.38 (1.17)	1.44 (1.21)	0.048
mstatus (%)			0.097
divorced	540 (23.6)	467 (20.4)	
married	479 (20.9)	554 (24.2)	
single	1272 (55.5)	1270 (55.4)	
age (mean (SD))	42.97 (11.29)	42.56 (10.80)	0.037
sex = Male (%)	1272 (55.5)	1340 (58.5)	0.060
edu (mean (SD))	11.70 (2.90)	11.80 (3.17)	0.034
vacation (mean (SD))	10.31 (2.80)	10.50 (2.66)	0.070
weight (mean (SD))	160.83 (42.83)	169.58 (44.10)	0.201
height (mean (SD))	64.65 (6.61)	64.91 (6.85)	0.038
hrfriend = Yes (%)	1220 (53.3)	1215 (53.0)	0.004
cxofriend = Yes (%)	1254 (54.7)	1318 (57.5)	0.056
insurance (%)			0.076
Covered	862 (37.6)	940 (41.0)	
Covered & Medicaid	398 (17.4)	393 (17.2)	
Covered & Medicare	10 (0.4)	10 (0.4)	
Medicaid	862 (37.6)	789 (34.4)	
Medicare	16 (0.7)	16 (0.7)	
Medicare & Medicaid	1 (0.0)	1 (0.0)	
Other	142 (6.2)	142 (6.2)	
flexspend = Yes (%)	770 (33.6)	957 (41.8)	0.169
retcont = Yes (%)	131 (5.7)	131 (5.7)	<0.001
race (%)			0.030
black	328 (14.3)	352 (15.4)	
other	148 (6.5)	148 (6.5)	
white	1815 (79.2)	1791 (78.2)	
disthome (mean (SD))	23.01 (8.04)	15.69 (8.68)	0.876
testscore (mean (SD))	63.40 (14.99)	56.57 (13.36)	0.481

Caliper = 3

The matching with caliper 1 doesn't work since it has a serious violation of the positivity assumption. It also only has 82 matching pairs. Caliper 3 will not work since the SMD for multiple covariates is unacceptably high which indicates the poor matching task. After tuning, we chose the caliper value = 2 as the result shown below. In this way, only minimal violation of the positivity assumption and most of the covariates would have little SMD to satisfy our requirement.

	Stratified by training		
	No	Yes	SMD
n	836	836	
manager = Yes (%)	106 (12.7)	106 (12.7)	<0.001
raise = Yes (%)	254 (30.4)	254 (30.4)	<0.001
salary (%)			<0.001
Under $20k	551 (65.9)	551 (65.9)	
$20-$40k	124 (14.8)	124 (14.8)	
$40-$80k	109 (13.0)	109 (13.0)	
> $80k	52 (6.2)	52 (6.2)	
children (mean (SD))	1.25 (1.06)	1.25 (1.08)	0.002
mstatus (%)			<0.001
divorced	166 (19.9)	166 (19.9)	
married	138 (16.5)	138 (16.5)	
single	532 (63.6)	532 (63.6)	
age (mean (SD))	44.52 (10.34)	43.87 (10.30)	0.063
sex = Male (%)	481 (57.5)	481 (57.5)	<0.001
edu (mean (SD))	11.47 (2.66)	11.54 (2.79)	0.029
vacation (mean (SD))	10.05 (2.51)	10.24 (2.46)	0.079
weight (mean (SD))	159.73 (37.66)	164.15 (36.25)	0.120
height (mean (SD))	64.52 (6.29)	64.74 (6.49)	0.034
hrfriend = Yes (%)	440 (52.6)	440 (52.6)	<0.001
cxofriend = Yes (%)	499 (59.7)	499 (59.7)	<0.001
insurance (%)			<0.001
Covered	265 (31.7)	265 (31.7)	
Covered & Medicaid	167 (20.0)	167 (20.0)	
Covered & Medicare	0 (0.0)	0 (0.0)	
Medicaid	387 (46.3)	387 (46.3)	
Medicare	1 (0.1)	1 (0.1)	
Medicare & Medicaid	0 (0.0)	0 (0.0)	
Other	16 (1.9)	16 (1.9)	
flexspend = Yes (%)	295 (35.3)	295 (35.3)	<0.001
retcont = Yes (%)	34 (4.1)	34 (4.1)	<0.001
race (%)			<0.001
black	87 (10.4)	87 (10.4)	
other	11 (1.3)	11 (1.3)	
white	738 (88.3)	738 (88.3)	
disthome (mean (SD))	22.94 (8.16)	17.69 (8.27)	0.640
testscore (mean (SD))	64.47 (14.11)	58.78 (11.86)	0.437

Caliper = 2

However, the SMD for `disthome` and `testscore` would not decrease below 0.1 for any caliper value. Although we theoretically would not classify them as confounders, we need to come out with further matching methods for comprehensive understanding.

4.2 McNemar Test

```
        McNemar's Chi-squared test with continuity correction

data:  indata
McNemar's chi-squared = 25.951, df = 1, p-value = 3.502e-07
```

Subsequent to the one-to-one matching process, we employed the McNemar Test to ascertain the effect of our treatment on the outcome within this model. The test results, which revealed a p-value of 3.502e-07, indicate that the treatment has a measurable impact on the outcome. Consequently, this supports the conclusion that, in this model, the observed changes in the outcome variable can be causally attributed to the treatment.

5 Propensity Score Matching

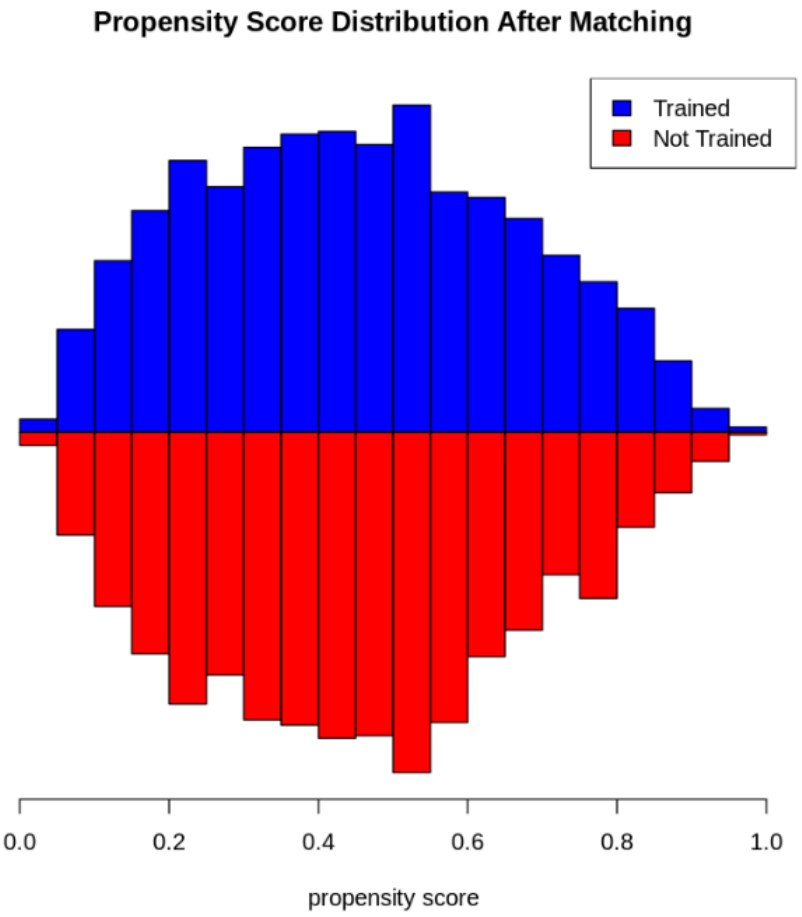
5.1Processing

Initially, logistic regression is employed to estimate the propensity scores for each individual in the study. These scores represent the probability of receiving the treatment, given their observed covariates.

With the propensity scores calculated, individuals in the treatment group are matched with those in the control group on a 1:1 basis, using a caliper of 0.1 to ensure that matched pairs have closely aligned propensity scores. This meticulous pairing seeks to create equivalently distributed covariates across treated and untreated units, effectively simulating a randomized experimental setting.

We evaluate the matches' quality by constructing tables and graphs to summarize the baseline characteristics of both groups. As we can see from the following table and graph, the matching provides enough cases for both sides, low SMD for each covariate, no violation of assumption, and symmetric distribution between trained and not trained. However, the distribution of propensity occurs in the situation of long tails.

	Stratified by training		
	No	Yes	SMD
n	1427	1427	
manager = Yes (%)	209 (14.6)	210 (14.7)	0.002
raise = Yes (%)	482 (33.8)	476 (33.4)	0.009
salary (%)			0.039
Under $20k	793 (55.6)	789 (55.3)	
$20-$40k	306 (21.4)	291 (20.4)	
$40-$80k	211 (14.8)	228 (16.0)	
> $80k	117 (8.2)	119 (8.3)	
children (mean (SD))	1.42 (1.20)	1.43 (1.20)	0.005
mstatus (%)			0.022
divorced	309 (21.7)	313 (21.9)	
married	315 (22.1)	326 (22.8)	
single	803 (56.3)	788 (55.2)	
age (mean (SD))	42.91 (11.77)	42.63 (11.04)	0.025
sex = Male (%)	792 (55.5)	793 (55.6)	0.001
edu (mean (SD))	11.71 (3.07)	11.74 (3.14)	0.008
vacation (mean (SD))	10.38 (2.89)	10.44 (2.63)	0.020
weight (mean (SD))	164.10 (47.24)	164.15 (41.52)	0.001
height (mean (SD))	64.69 (6.76)	64.79 (6.90)	0.016
hrfriend = Yes (%)	756 (53.0)	734 (51.4)	0.031
cxofriend = Yes (%)	811 (56.8)	816 (57.2)	0.007
insurance (%)			0.042
Covered	534 (37.4)	551 (38.6)	
Covered & Medicaid	252 (17.7)	262 (18.4)	
Covered & Medicare	8 (0.6)	7 (0.5)	
Medicaid	532 (37.3)	515 (36.1)	
Medicare	16 (1.1)	14 (1.0)	
Medicare & Medicaid	1 (0.1)	1 (0.1)	
Other	84 (5.9)	77 (5.4)	
flexspend = Yes (%)	542 (38.0)	531 (37.2)	0.016
retcont = Yes (%)	133 (9.3)	112 (7.8)	0.053
race (%)			0.017
black	218 (15.3)	227 (15.9)	
other	84 (5.9)	83 (5.8)	
white	1125 (78.8)	1117 (78.3)	
disthome (mean (SD))	19.72 (6.97)	19.85 (7.34)	0.017
testscore (mean (SD))	60.11 (15.40)	59.72 (12.16)	0.029



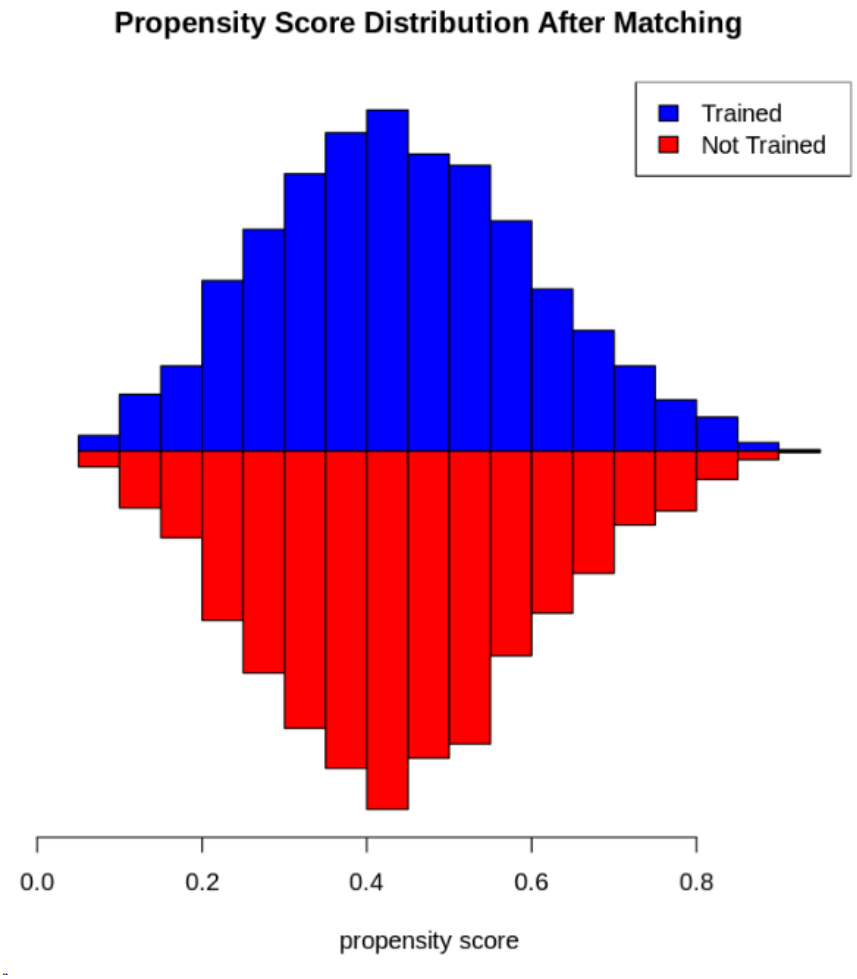
5.2 Checking the nature of ‘disthome’

`disthome` has been in the way of our matching tests as its (and only its) SMD remains high despite its high caliper. Thus, future study into the nature of ‘disthome’ was conducted:

1. X = `disthome`, which means distance from home to training center. This metric intuitively and statistically influence subjects’ compliance to training, thus X -> T
2. Whether X directly influences Y is furthered examined by checking if employees who are closer to the training center have a higher propensity of getting promoted, after controlling for T
    a. For T = ‘No’, `disthome` is not significant to Y with P-value = 0.38
    b. For T = ‘Yes’, `disthome` is marginally significant to Y with P-value = 0.03
    c. Our conclusion would be that there’s a very weak X ->Y
3. Given only X -> T -> Y without a strong X ->Y, it could be reasonable to exclude this X in matching not considering it as a confounder

After canceling out the `disthome` from the variables, the propensity score distribution with matching becomes more centralized which performs a better matching model.

	Stratified by training		
	No	Yes	SMD
n	1944	1944	
manager = Yes (%)	303 (15.6)	291 (15.0)	0.017
raise = Yes (%)	654 (33.6)	619 (31.8)	0.038
salary (%)			0.021
Under $20k	1070 (55.0)	1052 (54.1)	
$20-$40k	393 (20.2)	405 (20.8)	
$40-$80k	320 (16.5)	328 (16.9)	
> $80k	161 (8.3)	159 (8.2)	
children (mean (SD))	1.42 (1.21)	1.40 (1.19)	0.015
mstatus (%)			0.003
divorced	426 (21.9)	424 (21.8)	
married	442 (22.7)	443 (22.8)	
single	1076 (55.3)	1077 (55.4)	
age (mean (SD))	42.54 (11.53)	42.73 (10.87)	0.016
sex = Male (%)	1100 (56.6)	1104 (56.8)	0.004
edu (mean (SD))	11.69 (3.13)	11.73 (3.17)	0.012
vacation (mean (SD))	10.45 (2.94)	10.39 (2.60)	0.023
weight (mean (SD))	166.66 (48.04)	166.72 (42.59)	0.001
height (mean (SD))	64.77 (6.86)	64.76 (6.81)	0.001
hrfriend = Yes (%)	1014 (52.2)	1030 (53.0)	0.016
cxofriend = Yes (%)	1088 (56.0)	1105 (56.8)	0.018
insurance (%)			0.043
Covered	753 (38.7)	766 (39.4)	
Covered & Medicaid	339 (17.4)	346 (17.8)	
Covered & Medicare	12 (0.6)	10 (0.5)	
Medicaid	713 (36.7)	701 (36.1)	
Medicare	15 (0.8)	16 (0.8)	
Medicare & Medicaid	0 (0.0)	1 (0.1)	
Other	112 (5.8)	104 (5.3)	
flexspend = Yes (%)	733 (37.7)	749 (38.5)	0.017
retcont = Yes (%)	123 (6.3)	130 (6.7)	0.015
race (%)			0.013
black	303 (15.6)	303 (15.6)	
other	130 (6.7)	124 (6.4)	
white	1511 (77.7)	1517 (78.0)	
testscore (mean (SD))	58.70 (14.52)	58.87 (12.16)	0.013



5.3 McNemar Test

```
        McNemar's Chi-squared test with continuity correction

data:  ps_indata
McNemar's chi-squared = 225.58, df = 1, p-value < 2.2e-16
```

Following the propensity score matching, we conducted a McNemar Test to evaluate the causal impact of our treatment on the outcome variable. The obtained p-value, significantly lower than that of the one-to-one matching model, enhances our confidence in attributing a causal effect to the treatment. This finding solidifies our assertion that the treatment exerts a statistically significant influence on the outcome.

6. IPTW

6.1Processing

The application of Inverse Probability of Treatment Weighting (IPTW) in evaluating the impact of training programs on employees' test scores and distance from home yielded insightful results. Unlike other methods such as one-to-one Matching, Propensity Score Matching (PSM), and Instrumental Variable (IV) analysis, IPTW's distinctive approach to simulating random interventions by weighting outcomes presented a nuanced perspective on the significance of these factors.

	Stratified by training		
	No	Yes	SMD
n	5898.3	5555.1	
manager = Yes (%)	833.0 (14.1)	779.2 (14.0)	0.003
raise = Yes (%)	2150.4 (36.5)	1995.7 (35.9)	0.011
salary (%)			0.016
> $80k	438.2 (7.4)	433.7 (7.8)	
$20–$40k	1218.8 (20.7)	1138.7 (20.5)	
$40–$80k	924.7 (15.7)	855.9 (15.4)	
Under $20k	3316.7 (56.2)	3126.9 (56.3)	
children (mean (SD))	1.37 (1.18)	1.39 (1.19)	0.016
mstatus (%)			0.032
divorced	1322.5 (22.4)	1284.0 (23.1)	
married	1255.0 (21.3)	1230.3 (22.1)	
single	3320.8 (56.3)	3040.8 (54.7)	
age (mean (SD))	42.91 (11.85)	42.86 (11.16)	0.004
sex = Male (%)	3271.1 (55.5)	2973.4 (53.5)	0.039
edu (mean (SD))	11.64 (3.13)	11.69 (3.21)	0.017
vacation (mean (SD))	10.43 (2.91)	10.40 (2.58)	0.012
weight (mean (SD))	162.81 (48.52)	162.93 (42.19)	0.003
height (mean (SD))	64.73 (6.71)	64.53 (6.79)	0.029
hrfriend = Yes (%)	3111.1 (52.7)	2945.7 (53.0)	0.006
cxofriend = Yes (%)	3273.6 (55.5)	3112.7 (56.0)	0.011
insurance (%)			0.055
Covered	2119.7 (35.9)	2047.1 (36.9)	
Covered & Medicaid	1052.6 (17.8)	974.1 (17.5)	
Covered & Medicare	37.5 (0.6)	50.8 (0.9)	
Medicaid	2238.2 (37.9)	2075.3 (37.4)	
Medicare	101.4 (1.7)	104.9 (1.9)	
Medicare & Medicaid	14.9 (0.3)	5.5 (0.1)	
Other	334.0 (5.7)	297.4 (5.4)	
flexspend = Yes (%)	1980.0 (33.6)	1982.4 (35.7)	0.044
retcont = Yes (%)	718.5 (12.2)	692.1 (12.5)	0.008
race (%)			0.019
black	931.6 (15.8)	857.7 (15.4)	
other	362.6 (6.1)	364.2 (6.6)	
white	4604.2 (78.1)	4333.2 (78.0)	
disthome (mean (SD))	22.72 (8.77)	20.36 (8.35)	0.275
testscore (mean (SD))	61.89 (16.93)	60.58 (12.60)	0.088

The IPTW analysis revealed that the impact of training programs on test scores was less significant than observed with other methodologies. While one-to-one Matching and PSM showed a more pronounced effect, IPTW's findings suggest that when adjusting for confounders through weighting, the direct influence of training on improving test scores diminishes. This indicates that the observed benefits in test scores might be more sensitive to the choice of analysis method, underscoring the importance of considering multiple approaches in impact evaluations.

When compared to the treatment effects estimated through other methods, IPTW estimated a treatment effect of 1.36 with a significance level of <0.001 for the considered outcomes. This is in contrast to the stronger effects observed with 1v1 Matching (1.63, p=0.001), Propensity Score Matching (2.52, p<0.001), and Instrumental Variable analysis (1.26, p< 2.2e-16). The differences in estimated effects underscore the variability inherent in each method's approach to handling confounders and the distribution of covariates between treated and control groups.

7. Instrumental Variable

We select `disthome` (the distance from an employee's residence to the training location) as an Instrumental Variable (IV) is justified based on the fulfillment of the three core assumptions for a valid IV:

- **Relevance**: `disthome` is significantly associated with participation in the Career 2030 training program (the treatment variable). Theoretically, employees residing farther from the training venue may have lower participation rates due to logistical challenges. This indicates that `disthome` significantly influences the likelihood of training participation.
- **Exclusion**: The effect of `disthome` on the outcomes operates exclusively through its role in determining treatment status (i.e., whether an employee participates in the Career 2030 training). There are no unobserved variables that both influence `disthome` and directly affect the Promoted Status (the outcome variables), ensuring that the impact of `disthome` is solely mediated through its influence on training participation.
- **Exogeneity**: There is no direct causal path between `disthome` and the outcomes of interest (e.g., promotion or retention) except through its influence on training participation. This means that `disthome` should not directly affect an employee's promotion or retention outcomes independently of its effect on training participation.

7.1 stage1 model

```
Residual standard error: 0.3972 on 5971 degrees of freedom
Multiple R-squared:  0.3349,    Adjusted R-squared:  0.3318
F-statistic: 107.4 on 28 and 5971 DF,  p-value: < 2.2e-16
```

plot 7-1: all feature stage1 model's F-statistic

```
Residual standard error: 0.4246 on 5998 degrees of freedom
Multiple R-squared:  0.2364,    Adjusted R-squared:  0.2363
F-statistic:  1857 on 1 and 5998 DF,  p-value: < 2.2e-16
```

plot 7-2: single feature (disthome) stage1 model's F-statistic

In our two-fold stage 1 modeling approach, we first included all available features from the dataset, and then, in a separate model, we included only `disthome` as the variable of interest with `Training` being the dependent variable. In both models, the coefficient for `disthome` was -0.0238074 with a highly significant p-value of less than 2e-16.

The first model, incorporating all features, yielded an F-statistic of 107.4 with a p-value of less than 2.2e-16. The second model, with `disthome` as the sole independent variable, produced an even more pronounced F-statistic of 1875 with a p-value below 2.2e-16. These robust F-statistics from both models indicate a strong and significant relationship between `disthome` and the `Training` treatment, confirming the relevance of `disthome` as an instrumental variable.

The substantial F-statistic in the single-variable model cross-validates the relationship, mitigating concerns that the confidence in the multivariable model's F-statistic might be inflated due to overfitting and the conflation of effects from other covariates. This demonstrates that `disthome` is a strong predictor of training participation on its own, which supports its use as an instrumental variable in our analysis.

```
        Pearson's product-moment correlation

data:  tpd$disthome and tpd$training_num
t = -43.094, df = 5998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5053138 -0.4666646
sample estimates:
      cor
-0.486227
```

```
        Pearson's product-moment correlation

data:  tpd$disthome and tpd$promoted_num
t = -3.1853, df = 5998, p-value = 0.001453
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06632926 -0.01580677
sample estimates:
        cor
-0.04109428
```

Furthermore, we conducted Pearson's correlation tests to investigate the relationship between `disthome` and both `training` and `promoted`.

The correlation between `disthome` and `training_num` is significantly negative (correlation coefficient = -0.486227, p-value < 2.2e-16), which confirms the strong predictive power of `disthome` on training participation, satisfying the relevance condition for an instrumental variable.

The correlation between `disthome` and `promoted_num` is weak and not statistically significant (correlation coefficient = -0.0140928, p-value = 0.001453). This suggests that while `disthome` is related to the likelihood of participating in training, it does not

directly correlate with promotion, thus meeting the exclusion restriction for a valid instrument.

These findings corroborate the appropriateness of using `disthome` as an instrumental variable to identify the causal impact of training on promotion within the company.

7.2 stage 2 model

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.4271457  0.0717514  -5.953 2.78e-09 ***
training_est                   0.2302232  0.0238072   9.670  < 2e-16 ***
managerYes                     0.0182597  0.0224758   0.812 0.416586
raiseYes                      -0.0315479  0.0110316  -2.860 0.004254 **
salary.L                      -0.0485699  0.0229090  -2.120 0.034037 *
salary.Q                       0.0035409  0.0133156   0.266 0.790306
salary.C                       0.0019985  0.0126871   0.158 0.874839
children                      -0.0021752  0.0050491  -0.431 0.666631
mstatusmarried                -0.0843966  0.0162252  -5.202 2.04e-07 ***
mstatussingle                 -0.1450840  0.0127930 -11.341  < 2e-16 ***
age                            0.0054685  0.0005074  10.777  < 2e-16 ***
sexMale                        0.0378397  0.0107273   3.527 0.000423 ***
edu                            0.0024941  0.0017975   1.388 0.165326
vacation                      -0.0177732  0.0018697  -9.506  < 2e-16 ***
weight                        -0.0005582  0.0001176  -4.748 2.10e-06 ***
height                        -0.0004744  0.0007896  -0.601 0.547974
hrfriendYes                    0.0569985  0.0098753   5.772 8.24e-09 ***
cxofriendYes                   0.1197707  0.0100923  11.868  < 2e-16 ***
insuranceCovered & Medicaid   -0.0187576  0.0178517  -1.051 0.293416
insuranceCovered & Medicare    0.0246098  0.0644707   0.382 0.702682
insuranceMedicaid             -0.0061119  0.0148321  -0.412 0.680302
insuranceMedicare             -0.0231498  0.0418011  -0.554 0.579730
insuranceMedicare & Medicaid  -0.1806385  0.1003195  -1.801 0.071811 .
insuranceOther                 0.0239566  0.0236062   1.015 0.310221
flexspendYes                  -0.0155606  0.0113152  -1.375 0.169124
retcontYes                     0.0357561  0.0165404   2.162 0.030678 *
raceother                     -0.0011690  0.0234836  -0.050 0.960300
racewhite                     -0.0167692  0.0139751  -1.200 0.230212
testscore                      0.0164314  0.0003904  42.089  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3798 on 5971 degrees of freedom
Multiple R-squared:  0.3686,    Adjusted R-squared:  0.3656
F-statistic: 124.5 on 28 and 5971 DF,  p-value: < 2.2e-16
```

In the second stage of our model, we utilized the predicted values of training participation (`training_est`) from the first stage model as the explanatory variable to assess its impact on promotion outcomes (`promoted_num`). This model incorporates all covariates along with the predicted training variable.

The results show that the estimated coefficient for `training_est` is positive (0.2302232) and highly statistically significant (p-value < 2e-16), indicating a substantial positive correlation between the estimated value of training and promotion outcomes. The exponentiated coefficient (1.258881) suggests that, ceteris paribus, participation in training is estimated to increase the probability of promotion by approximately 25.89%.

The overall fit of the model is robust, with an F-statistic of 124.5 (on 28 and 5971 degrees of freedom, p-value < 2.2e-16), indicating that at least one predictor in the model has a significant impact on promotion outcomes.

8. Recommendation

8.1 Comparing models

	Theoretical Pros	**Theoretical Cons**	**Bias**
1v1 Matching	Straightforward to implement. Can be tailored by tuning Caliper	Does not account for unobserved confounders Sensitive to the choice of matching algorithm and distance metric Assumes sufficient overlap in the covariate distributions between treated and control	The remaining SMD Sensitivity caliper = 2
Propensity Score Matching	Handle multiple covariates simultaneously. Efficient when there are many covariates Can be tailored by tuning Caliper	Requires estimation of propensity scores Assumes sufficient overlap in the covariate distributions between treated and control	Difference in Propensity Score distribution PSM Sensitivity caliper = 0.1
IPTW	Simulates random interventions by weighting the outcomes	Prone to instability: sensitive to extreme weights and the propensity score model	
Instrumental Variable	Can address unobserved confounding Robust to model misspecification	Requires the identification of a valid instrument: assumes the instrument only affects the outcome through its effect on the treatment	IV not perfectly correlated with T Use F-statistic to determine

	Treatment Effect	P-Value
1v1 Matching	1.63	<0.001
Propensity Score Matching	2.43	<0.001
IPTW	1.36	<0.001
Instrumental Variable	1.26	< 2.2e-16

In addition to the theoretical pros and cons above, the models' practical performance compares in the following ways:

- One-to-one matching would contain high bias as the SMD for `disthome` and `testscore` would not decrease below 0.1 for any caliper value
- We concluded the nature of X=`disthome` in 5.2 to be strongly X->T->Y and weakly X->Y, which means
    - Propensity score and IPTW Matching still yielding high SMD of `disthome` will lead to bias but is not too worrying since its effect on Y is weak
    - Assumptions for `disthome` to be an instrumental variable is generally satisfied with minimal violation cause by the weak X->Y remains
- All models have a significant P-value while the instrumental variable one is the most confident in its coefficient.
- When evaluating consistency, it's not just about statistical significance but also about the magnitude and direction of the estimated treatment effects. If the estimated treatment effects from different methods are similar in magnitude and direction, they cross-validate each other, which is a pattern we see between one-to-one Matching, IPTW and instrumental variables' results.

Considering all, we have chosen the IV (Instrumental variable) approach as our primary basis for determining causal effects for the following reasons: after one-to-one matching and IPTW, some covariates exhibited a Standardized Mean Difference (SMD) greater than 0.2, indicating significant bias. Although the p-value are significant in the PSM analysis, the propensity score distribution of PSM with `disthome` feature was too dispersed, almost spanning the full range from 0 to 1. Ideally, we would expect a more concentrated distribution, particularly between 0.2 and 0.8, as a broad distribution suggests potential imbalance in covariates post-matching. Additionally, while the P-value from PSM indicates statistical significance, it remains lower than the P-value from IV. Although we typically cannot directly compare P-values across different models, a smaller P-value here can be interpreted as stronger statistical evidence, suggesting that the results from the IV method are more reliable.

Regarding the assumptions for IV, `disthome` satisfied the three main criteria required for an effective instrument: Relevance, Exclusion, and Exogeneity. We conducted a series of tests to validate its effectiveness, including the Hausman test for instrument validity, Pearson correlation to establish the relationship between the instrument and the treatment, as well as first and second-stage regression analyses, and tests for weak instruments. All these tests indicated that `disthome` is a strong and valid instrument. Particularly, the F-statistic in the second-stage model was substantially high, suggesting that the bias from using `disthome` as an IV is minimal, reinforcing the likelihood that the Exogeneity assumption—which is difficult to test statistically—is likely valid. Therefore, `disthome` is well-suited to estimate the causal effect of the Career 2030 training program.

8.2 Business Impact

With Instrumental variable, we can conclude that the odds of being promoted among individuals who received training (T=1) are 1.26 times higher than the odds of being promoted among individuals who did not receive training (T=0), coincide with the expectation – “training program aimed at fostering the career development of its workforce”.

Considering the cost rendered by treatment corruption in this experiment, there’s several recommendation for executing similar RCT going forward:

- **Strengthen Compliance Monitoring:** Implement rigorous compliance monitoring mechanisms to ensure that employees selected for training actually comply with program requirements. This can include regular check-ins, progress tracking, and consequences for non-compliance.
- **Address Attrition Bias:** Account for attrition rates between treatment groups by collecting and analyzing data on employee turnover during the trial period. Adjust for attrition imbalances using statistical methods if occurs.
- **Mitigate Managerial Intervention:** Implement protocols and safeguards to prevent managerial intervention in the randomization process. Ensure that randomization procedures are transparent, independent, and free from managerial influence. Consider using centralized randomization systems or blinded allocation procedures to reduce the potential for bias.

Appendix

```
Training Group: No
    children          age             edu           vacation
 Min.   :0.000   Min.   :18.00   Min.   :-1.00   Min.   : 6.00
 1st Qu.:1.000   1st Qu.:35.00   1st Qu.:10.00   1st Qu.: 8.00
 Median :1.000   Median :45.00   Median :12.00   Median :10.00
 Mean   :1.348   Mean   :43.26   Mean   :11.51   Mean   :10.31
 3rd Qu.:2.000   3rd Qu.:52.00   3rd Qu.:13.00   3rd Qu.:12.00
 Max.   :5.000   Max.   :71.00   Max.   :30.00   Max.   :18.00
     weight           height         disthome       testscore
 Min.   : 44.72   Min.   :52.96   Min.   :11.00   Min.   : 21.00
 1st Qu.:127.16   1st Qu.:59.01   1st Qu.:18.00   1st Qu.: 56.00
 Median :153.29   Median :63.55   Median :26.00   Median : 67.50
 Mean   :158.41   Mean   :64.48   Mean   :25.62   Mean   : 66.21
 3rd Qu.:182.10   3rd Qu.:69.09   3rd Qu.:33.00   3rd Qu.: 77.50
 Max.   :450.12   Max.   :81.64   Max.   :40.00   Max.   :100.00
Training Group: Yes
    children         age             edu          vacation        weight
 Min.   :0.00   Min.   :18.00   Min.   : 0.0   Min.   : 6.0   Min.   : 71.4
 1st Qu.:1.00   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 8.0   1st Qu.:139.7
 Median :1.00   Median :44.00   Median :12.0   Median :10.0   Median :165.1
 Mean   :1.44   Mean   :42.56   Mean   :11.8   Mean   :10.5   Mean   :169.6
 3rd Qu.:2.00   3rd Qu.:51.00   3rd Qu.:14.0   3rd Qu.:12.0   3rd Qu.:192.5
 Max.   :5.00   Max.   :70.00   Max.   :27.0   Max.   :18.0   Max.   :536.1
     height         disthome       testscore
 Min.   :52.96   Min.   : 1.00   Min.   :17.00
 1st Qu.:59.45   1st Qu.: 8.00   1st Qu.:48.50
 Median :64.27   Median :16.00   Median :58.00
 Mean   :64.91   Mean   :15.69   Mean   :56.57
 3rd Qu.:69.61   3rd Qu.:23.00   3rd Qu.:66.50
 Max.   :81.64   Max.   :30.00   Max.   :85.00
```

```
promoted partitioned by training:

Training: No

       No       Yes
0.3680237 0.6319763

Training: Yes

       No       Yes
0.3190746 0.6809254
```

```
manager partitioned by training:

Training: No

      No      Yes
0.869237 0.130763

Training: Yes

       No       Yes
0.8345701 0.1654299
```

```
salary partitioned by training:

Training: No

    > $80k   $20-$40k   $40-$80k Under $20k
0.07252629 0.20032354 0.14316527 0.58398490

Training: Yes

    > $80k   $20-$40k   $40-$80k Under $20k
0.08904409 0.20558708 0.17983413 0.52553470
```

```
mstatus partitioned by training:

Training: No

 divorced   married    single
0.2453492 0.1952009 0.5594500

Training: Yes

 divorced   married    single
0.2038411 0.2418158 0.5543431
```

```
sex partitioned by training:

Training: No

   Female      Male
0.4613103 0.5386897

Training: Yes

   Female      Male
0.4151026 0.5848974
```

```
hrfriend partitioned by training:

Training: No

       No       Yes
0.4605015 0.5394985

Training: Yes

       No       Yes
0.4696639 0.5303361
```

```
cxofriend partitioned by training:

Training: No

       No       Yes
0.4648153 0.5351847

Training: Yes

       No       Yes
0.4247054 0.5752946
```

```
insurance partitioned by training:

Training: No

          Covered  Covered & Medicaid  Covered & Medicare           Medicaid
        0.333243462         0.172553249         0.007279590         0.409005123
   Medicare Medicare & Medicaid               Other
        0.022917228         0.003774602         0.051226746

Training: Yes

          Covered  Covered & Medicaid  Covered & Medicare           Medicaid
        0.4103011785        0.1715408119        0.0043649062        0.3443910956
   Medicare Medicare & Medicaid               Other
        0.0069838498        0.0004364906        0.0619816674
```

```
flexspend partitioned by training:

Training: No

       No       Yes
0.7150175 0.2849825

Training: Yes

       No       Yes
0.5822785 0.4177215
```

```
retcont partitioned by training:

Training: No

      No      Yes
0.839849 0.160151

Training: Yes

        No        Yes
0.94281973 0.05718027
```

```
race partitioned by training:

Training: No

     black      other      white
0.16743057 0.05958479 0.77298463

Training: Yes

     black      other      white
0.15364470 0.06460061 0.78175469
```