SC1015 Mini Project

Kynan Lau, Ong Yu Jing, Yap Xin Yi

Table of Contents

Problem Definition & Motivation

- A major problem in workplaces now are the unhappiness of employees due to the unsuitability of the work environment to their personal preferences

- Given ratings of the individual metrics, predict overall rating

- Problem: improve employee retention

Dataset

Glassdoor Job Reviews

A large dataset of job reviews with textual features and numerical targets

Size : 838566

Number of columns: 18

- This large dataset contains job descriptions and rankings among various criteria such as work-life balance, income, culture, etc.
- Each metrics in the data set led to the determination of overall rating of their job at a certain company

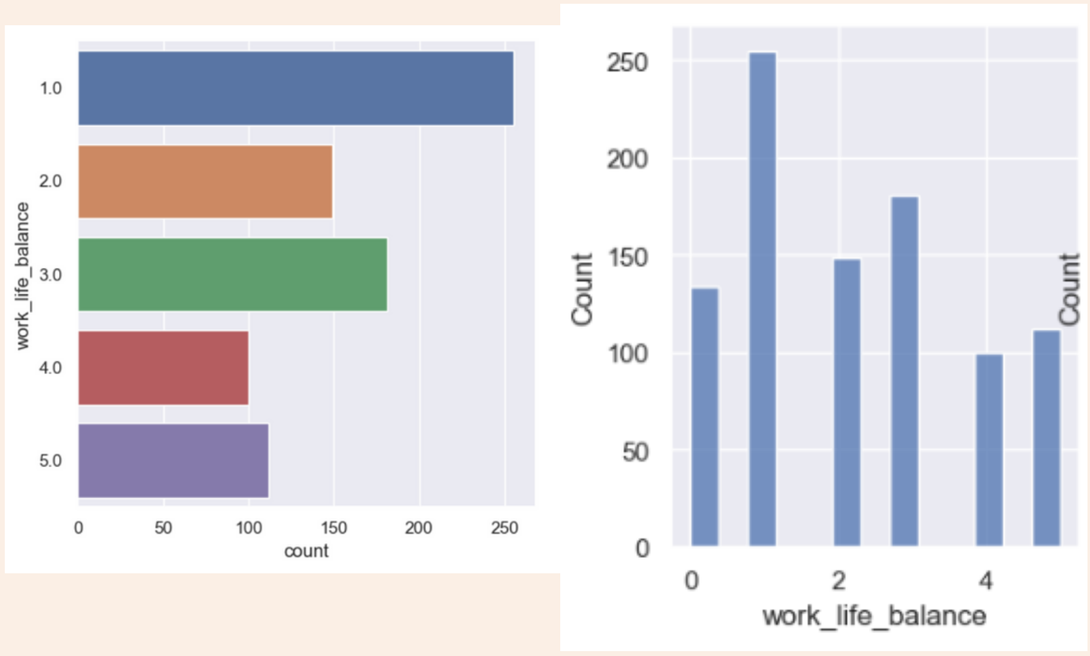
Data Cleaning & Preparation

- **Extracting data for the company 'Aldi'** : Aldi is a discount grocery supermarket in over 20 countries (12,419 stores worldwide) so it would have a wider reach globally
- **Removed columns:**

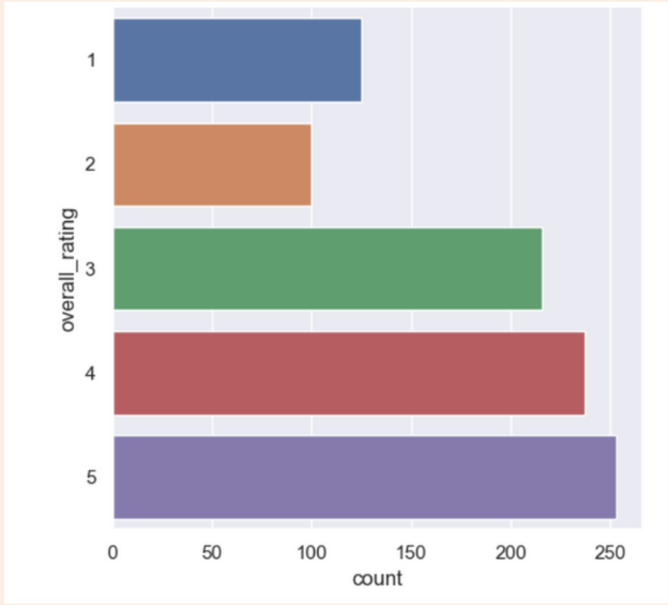
Column(s) Removed	**Reason**
'headline', 'date_review'	not relevant to our regression
'pros', 'cons'	not able to find out the correlation of the comments given to the overall rating hence we will ignore these columns and focus on those with numerical or categorical ratings
'diversity_inclusion'	most of the reviews given left this column blank hence we decided to remove this column as the accuracy of the prediction of the overall_rating might be affected due to the low sample size

Visualization of Data

Frequency of each rating corresponding to each metric:
using histplot and catplot

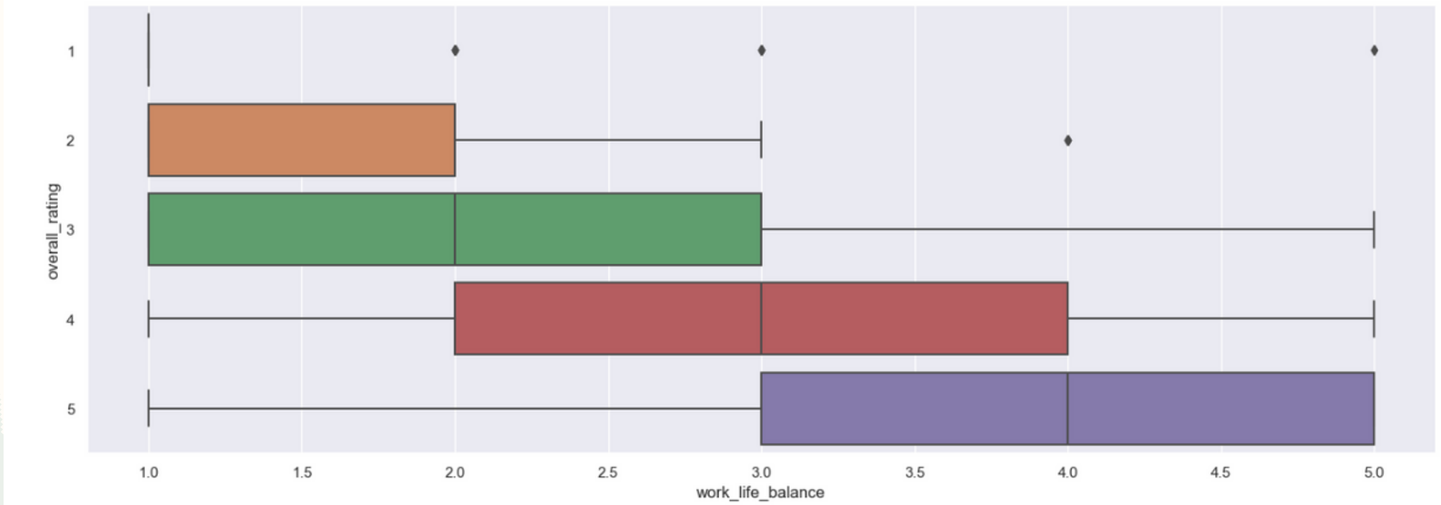


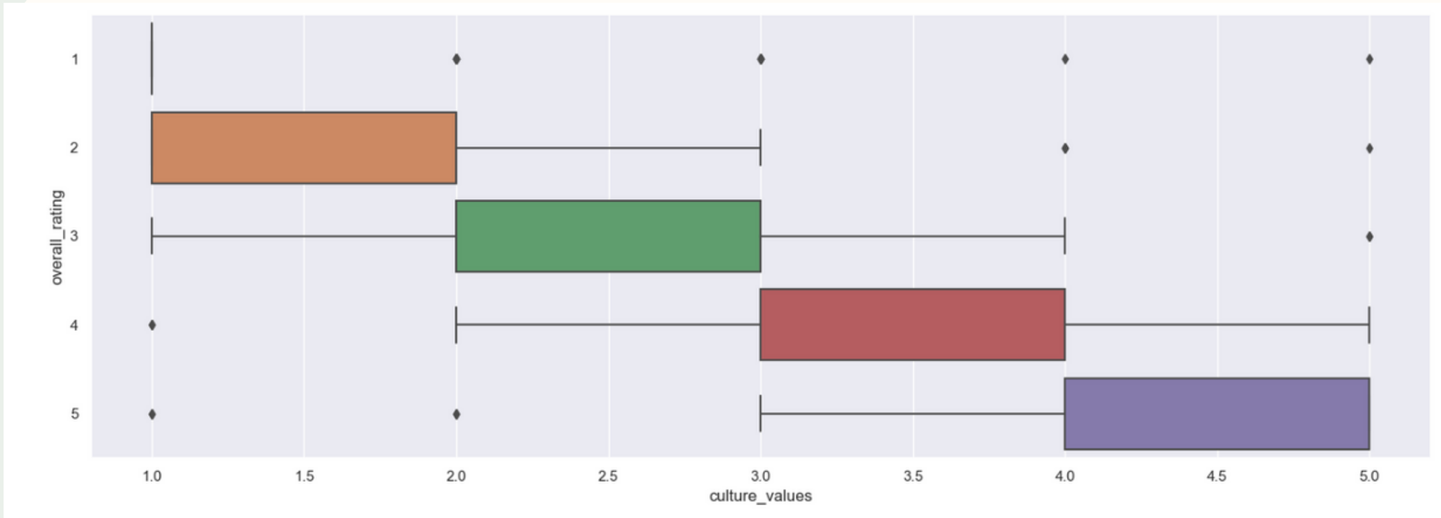
Frequency of overall_rating:
using catplot



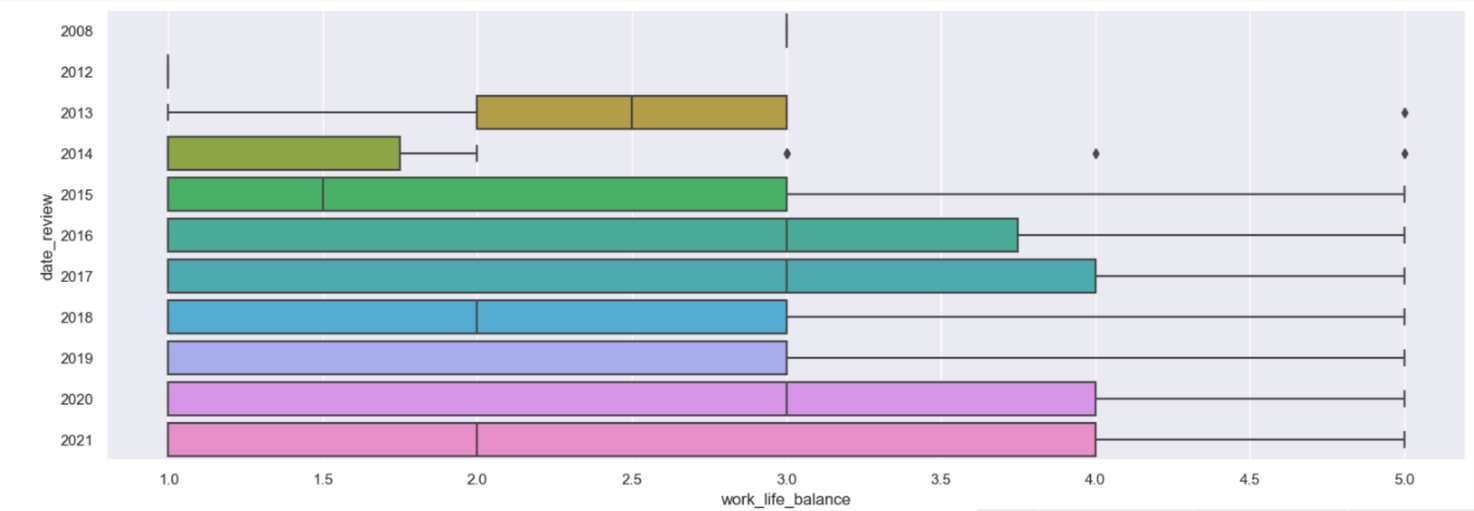
Exploratory Data Analysis

Compare overall_rating against each metric individually to see how they each affect overall_rating using boxplot





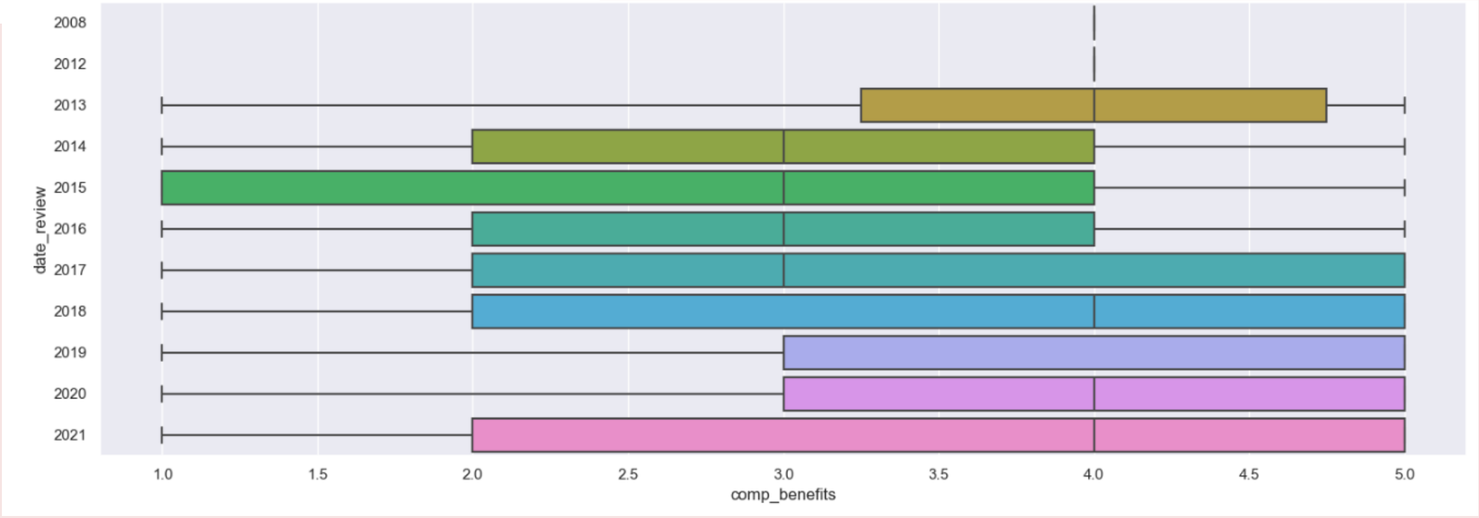
Exploratory Data Analysis



See frequency of each metric against time using boxplot

- Comparing through the years, work_life_balance has been consistently the worst
- comp_benefits has been on the rise

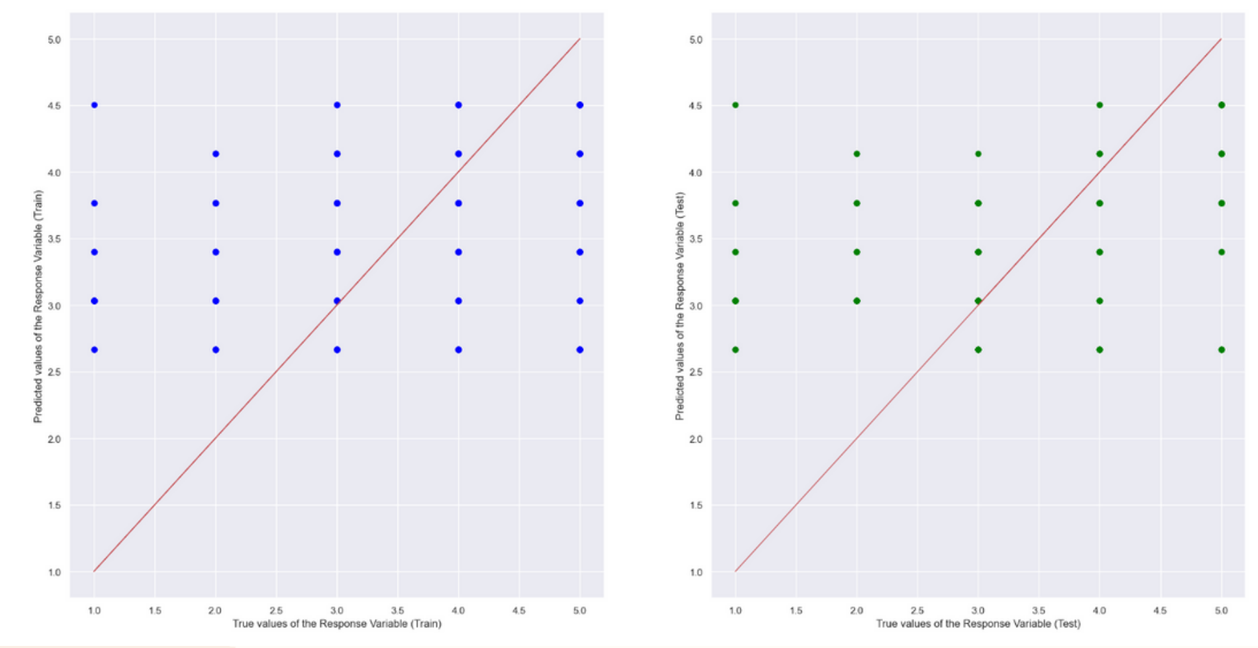
- culture_values, career_opp and senior_mgmt has been average and stagnant (graphs not shown)

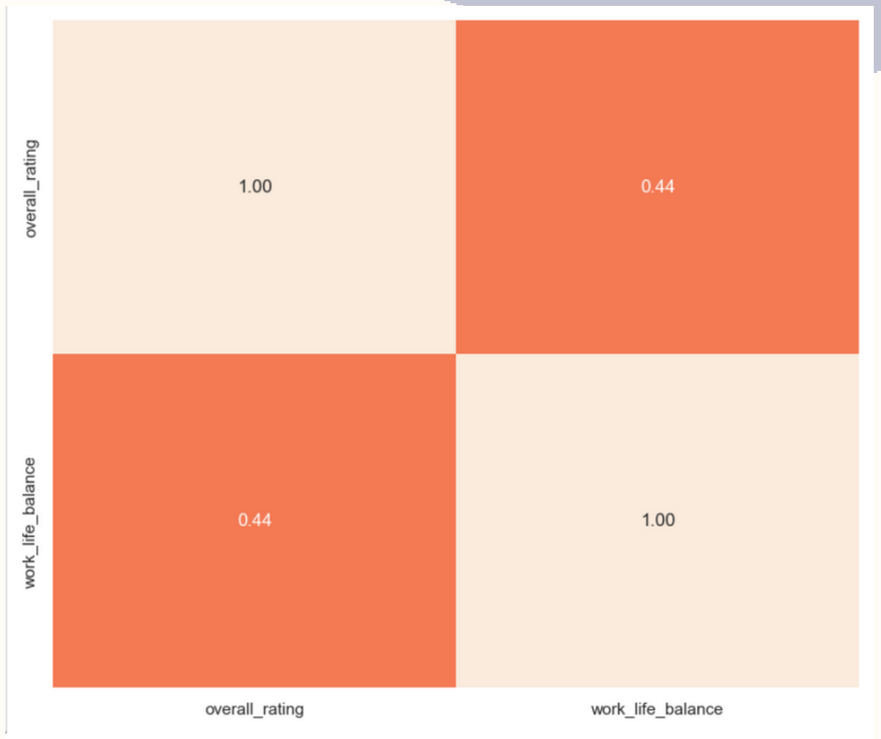


Linear Regression

Uni-variate

Response Variable : Overall Rating
Predictor Feature : Work life balance
Regression Model : Overall rating = a x work_life_balance + b





Findings:

1. Correlation between the 2 variables is 0.44
2.

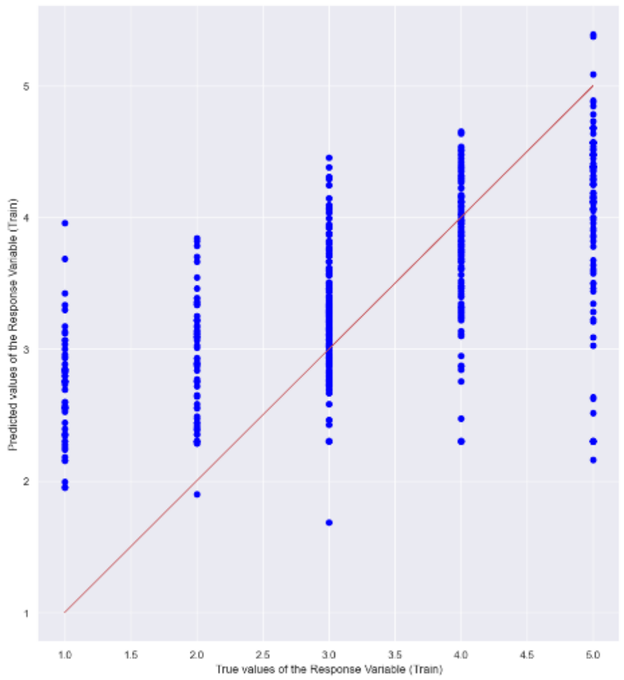
```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.19576707223876222
Mean Squared Error (MSE)          : 1.435413583821403

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.25116139778560276
Mean Squared Error (MSE)          : 1.3861700619081885
```

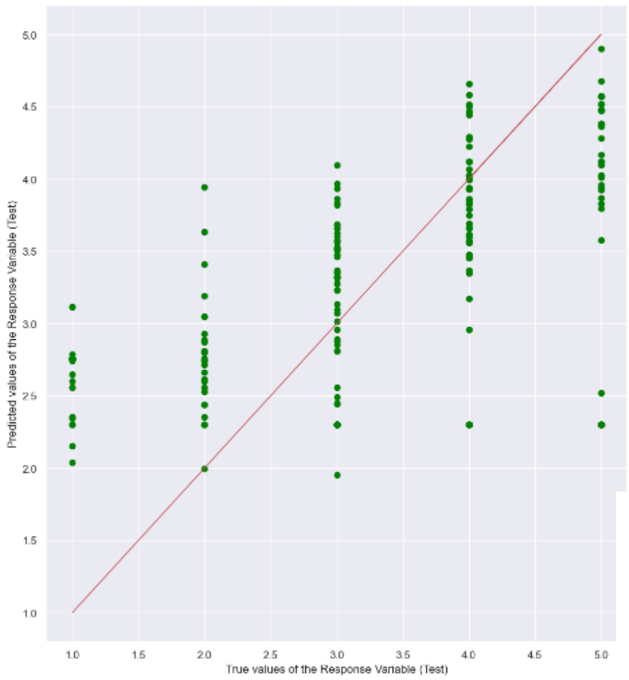
Multi-variate

Response Variable : Overall Rating

Predictor Feature : ALL the numerical metrics

Regression Model : Overall rating = a x metrics_df + b

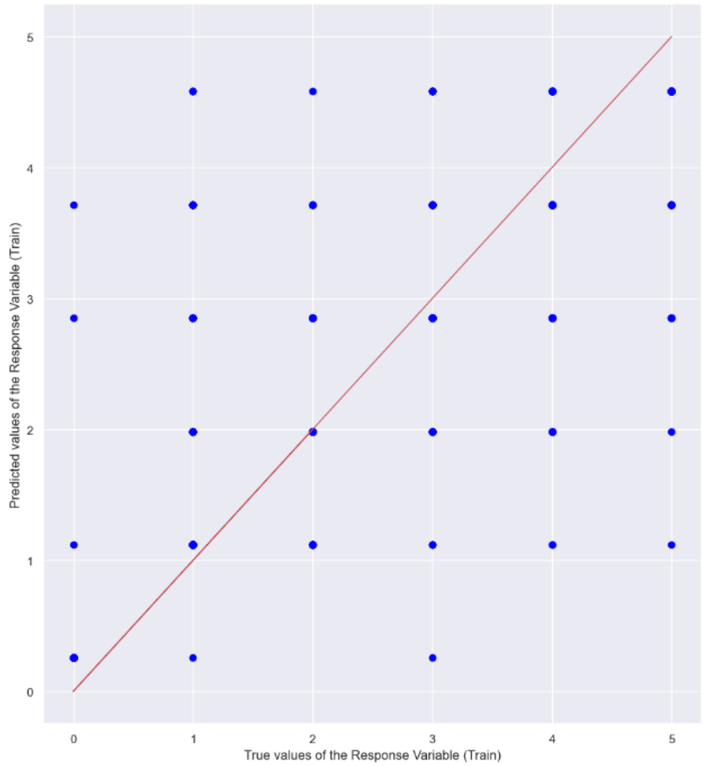


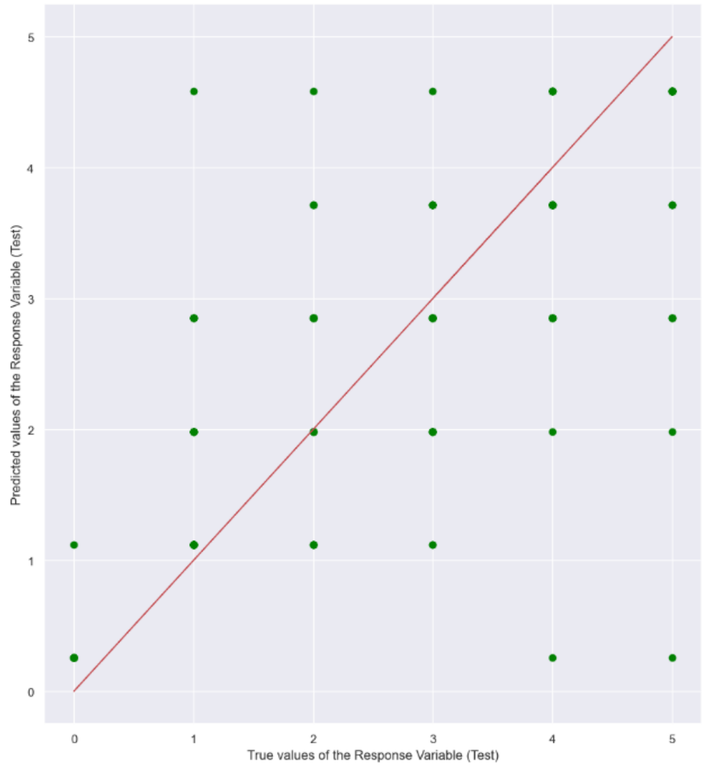


Findings:

1. culture_values has the highest correlation to overall_rating
2.
```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.34032897575154797
Mean Squared Error (MSE)          : 1.2072002490836613

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.30626076973002536
Mean Squared Error (MSE)          : 1.2075808591153383
```

Bi-variate Relationship (between senior_mgmt and culture_values)



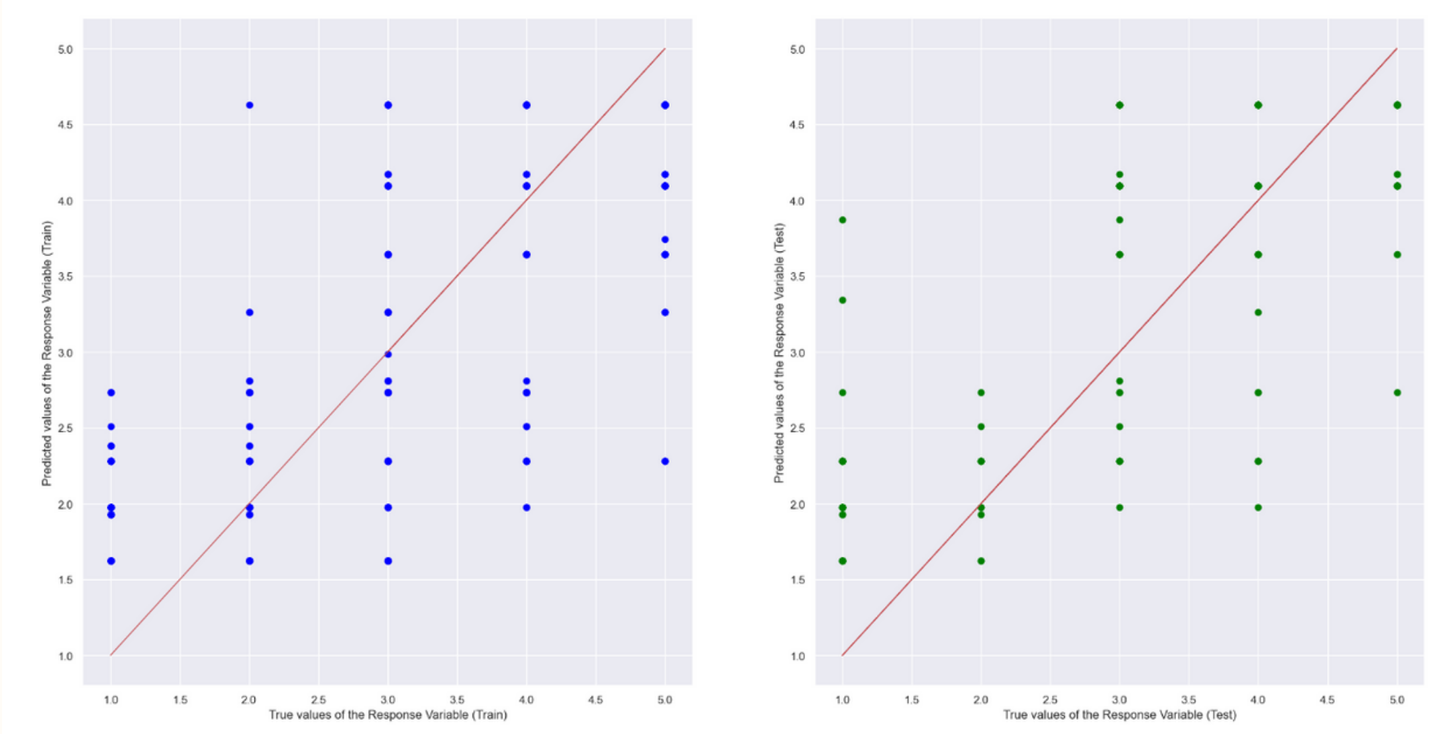


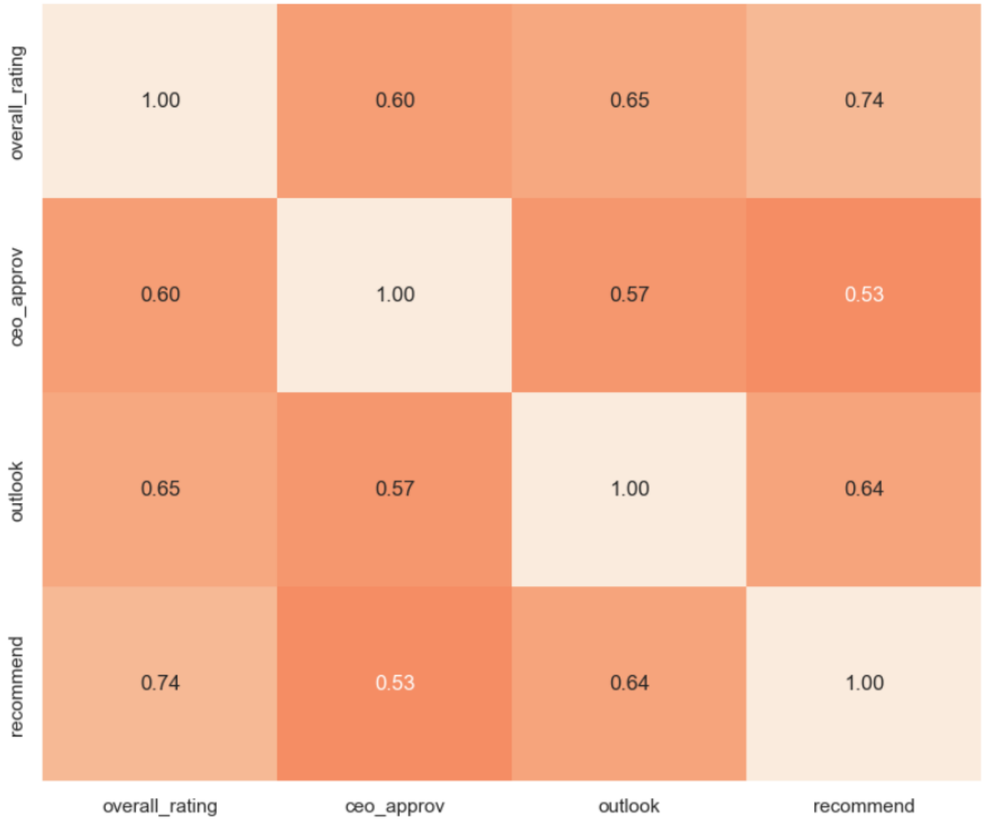

```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.7526788467837365
Mean Squared Error (MSE)        : 0.7354291990840173

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.7168568074880879
Mean Squared Error (MSE)        : 0.821573698974589
```

Findings:

1. Much higher explained variance and lower mean squared error -> much higher predictive power in this model and it has a lower margin of error
2. ALDI should aim to have both higher culture values and better senior management, so there is higher chance of improving overall rating, and possibly employee retention in ALDI

Using Categorical Variables to Predict overall_rating






```
Goodness of Fit of Model          Train Dataset
Explained Variance (R^2)          : 0.632512546640469
Mean Squared Error (MSE)          : 0.7294997647826902

Goodness of Fit of Model          Test Dataset
Explained Variance (R^2)          : 0.6007742542773833
Mean Squared Error (MSE)          : 0.7984322386604279
```

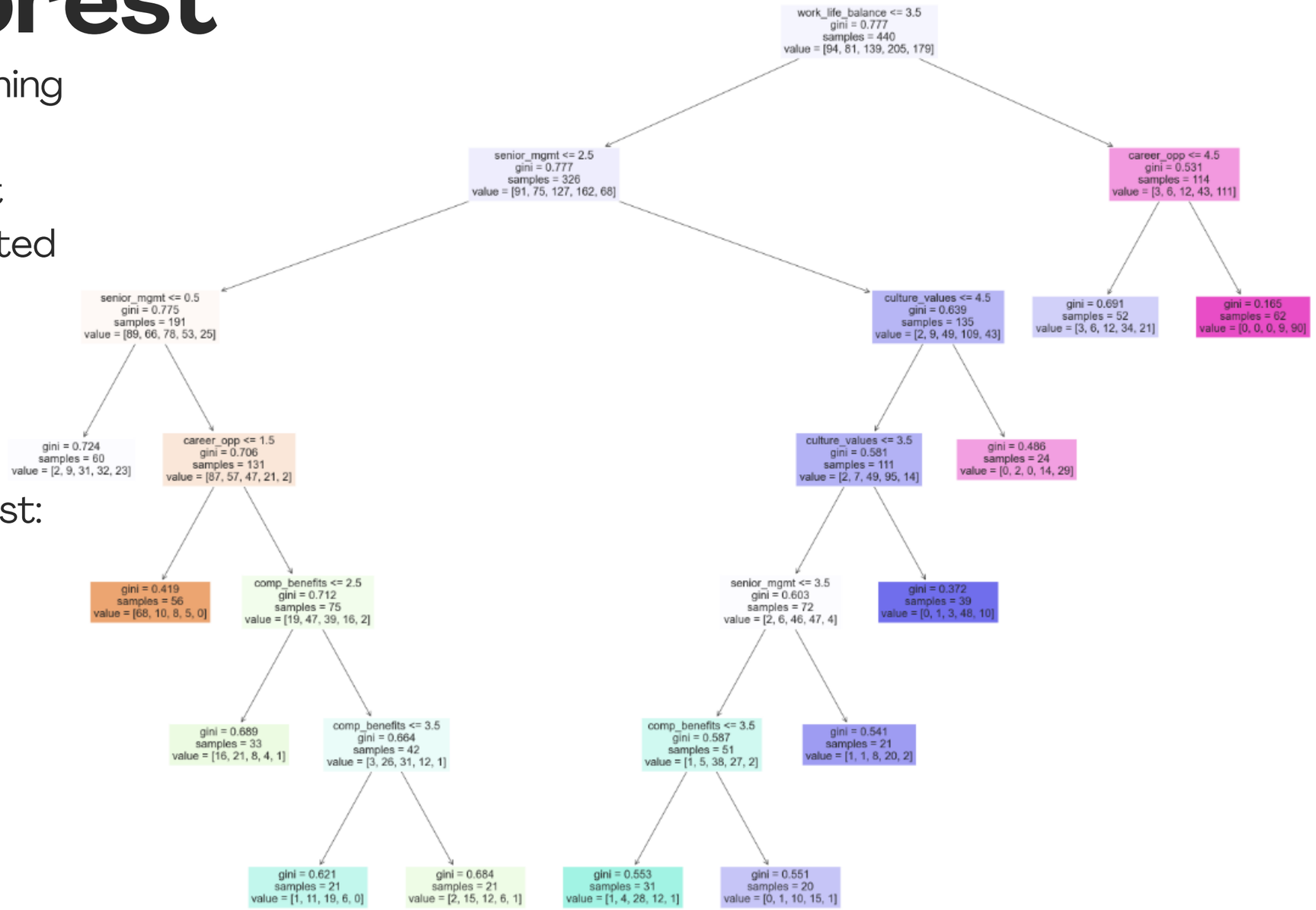
Findings:

recommend has the highest correlation to overall_rating

Random Forest

- Used Hyperparameter tuning using randomized search
  - To determine the best outcome for the created model, random combinations of the hyperparameters are applied.
- Accuracy of random forest: 51.9%
-

	Metric	Imp
1	culture_values	0.378802
4	senior_mgmt	0.279097
2	career_opp	0.210359
0	work_life_balance	0.081930
3	comp_benefits	0.049813



Logistic Regression

- Determine level of influence of categorical variables on overall_rating
- Convert each variable into numerical variables by letter_to_num
- Accuracy: 0.51 (train), 0.44 (test)

Logistic Regression

- Convert each variable into numerical variables by get_dummies
- Accuracy: 0.51 (train), 0.42 (test)

	overall_rating	recommend_v	recommend_x	ceo_approv_r	ceo_approv_v	ceo_approv_x	outlook_r	outlook_v	outlook_x
178	1	0	1	1	0	0	1	0	0
179	3	0	1	1	0	0	1	0	0
180	4	1	0	0	1	0	0	1	0
181	5	1	0	0	1	0	0	1	0
182	5	1	0	0	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...
1100	1	0	1	0	0	1	1	0	0
1102	1	0	1	0	0	1	1	0	0
1104	5	1	0	0	1	0	0	1	0
1105	5	1	0	0	1	0	0	1	0
1107	2	0	1	0	0	1	0	0	1

However, this shows that the machine learning does not work as well on this categorical variables:
Factor 1: Glassdoor separating it using symbols that are not detailed, lead to results that are less than ideal to predict
Factor 2: The low accuracy on train and test set has shown that the variables itself are limited, and not the fault of the machine being trained

Findings and Conclusion

1. Random Forest

With further resampling, we can better predict the ratings using numerical variables. The results align with the ones in linear regression.

2. Logistic Regression

Not recommended using only categorical variables to predict overall_rating

Data Driven Insights

- Maintain strong culture values and capable senior management together to continually improve ratings while possibly improving employee retention.
- Company benefits increased over the years, whereas every other metric remained average and stagnant. ALDI employees has poor work-life-balance.
- In terms of categorical variables, perhaps ALDI should look to improve outlook, which employees neglect and deem not important

THANK YOU!