2025

DATDRD05-T07 - Introduction to modelling Final assignment

2024-2025

HAKIM SHAIBU

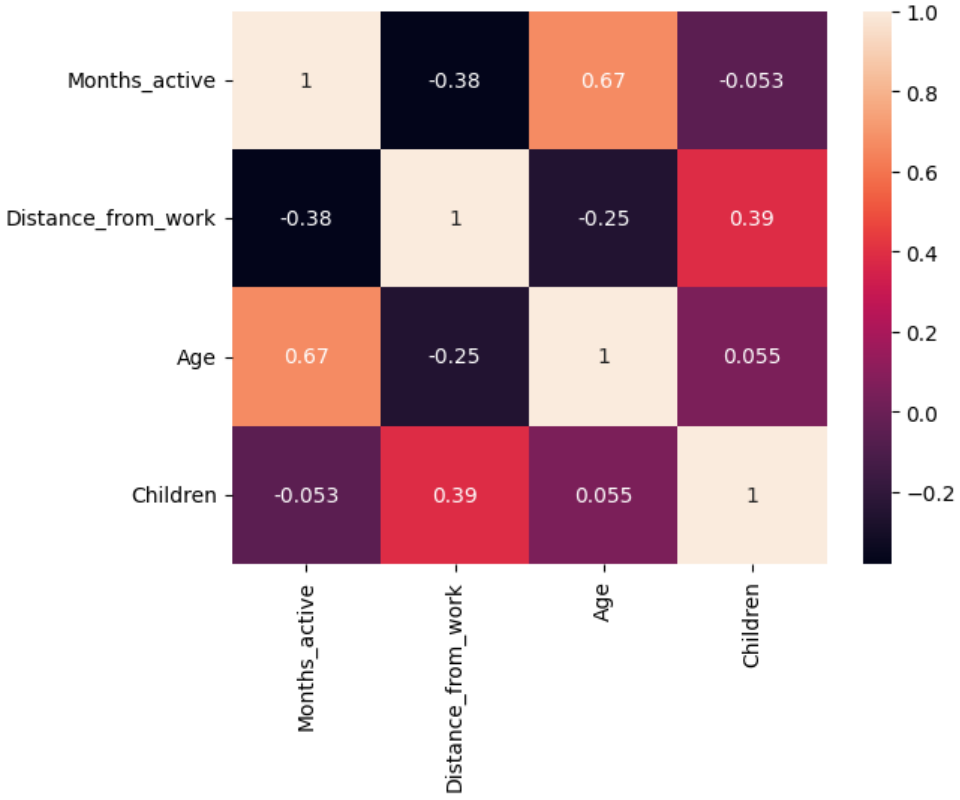
Hakim Shaibu -1641405

Lecturer: Tijmen Weber

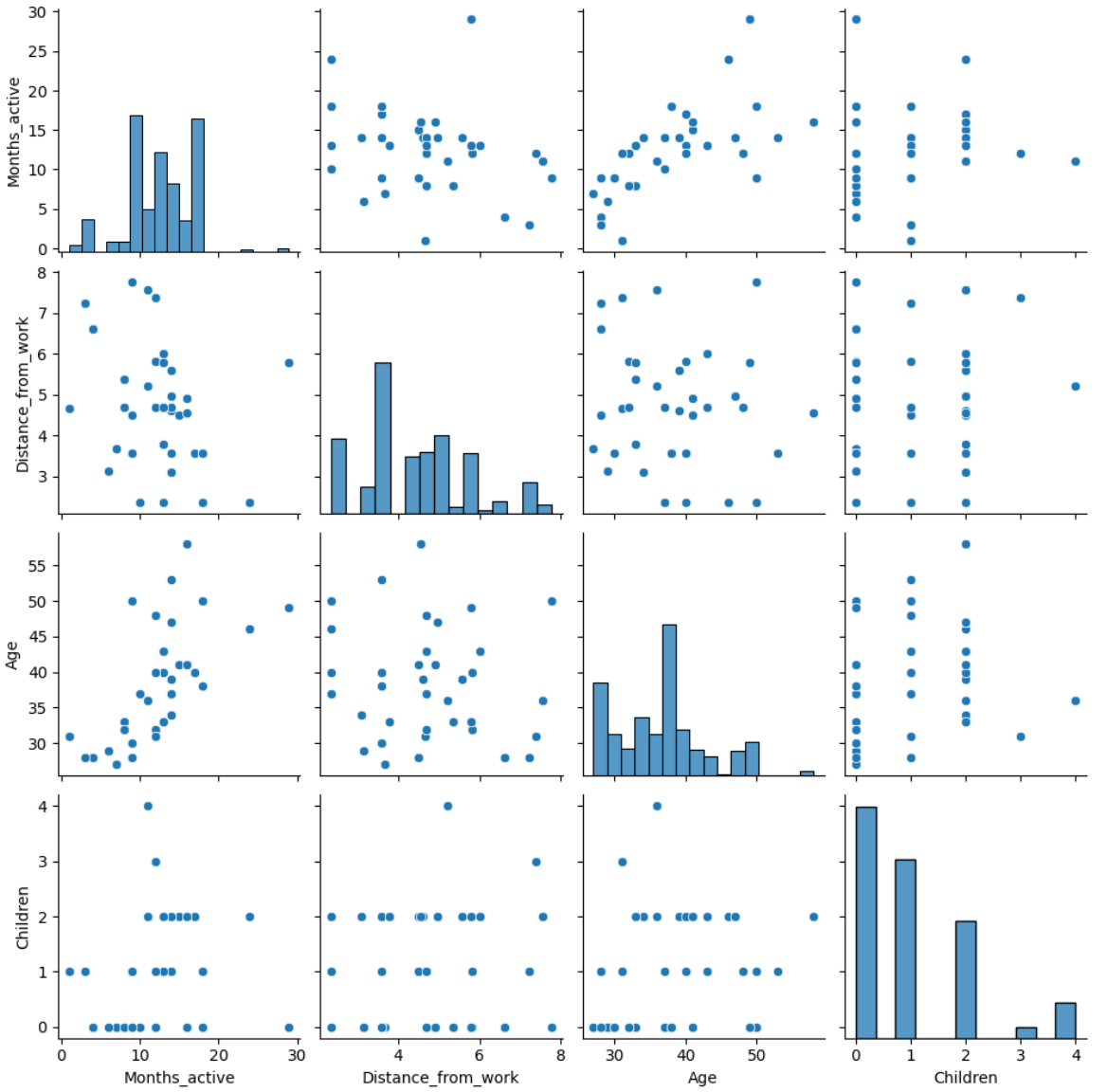
Contents

Link to GitHub repository

Assignment 1.1



Assignment 1.2



Assignment 1.3

To answer management's question about if there are correlation between the variables interested in, a correlation heatmap and scatterplot matrix were created. This will enable us to visualise and conclude on which pairs of variables are significantly correlated.

Correlation heatmap

The correlation heatmap visualizes the relationships between variables on a scale from -1 to 1, displayed in a color-coded matrix. A correlation close to 1 indicates a strong positive relationship meaning as one variable increases, the other tends to increase as well. A correlation near -1 signifies a strong negative relationship that is when one variable rises, the other tends to decrease. A correlation around 0 means there is little to no linear relationship between the variables.

A correlation of 1 appears when a variable is compared with itself, for example, 'Age' vs 'Age' or 'Months active' vs 'Months active'. This is expected because any variable is perfectly correlated with itself. It just means there's a 100% match, and it's a normal part of the heatmap's diagonal.

Therefore, from the resulting heatmap the strongest positive relationship appears to be the age of an employee and the months active with a correlation of 0.67. This means that the older the employee the more likely they are to stay longer with the company.

The Strongest negative correlation is between 'distance from work' and 'months active' with a correlation of -0.37. This means that employees who travel longer to get to work tend to stay with the company for a shorter period.

Scatterplot Matrix

The scatterplot matrix allows for a more visual way to explore the relationships between different variables. Each small plot shows how two variables relate to each other, and the diagonal shows the distribution of each individual variable.

The trend in the spread of the variables is directly related to the correlation heatmap. Over here:

- If the dots are in an upward trend, there is a positive correlation.
- If the dots are in a downward trend, there is a negative correlation.
- If the dots are randomly scattered, there is little to no relation.
- The tighter the dots are together the stronger the relationship.

For example, the plot comparing Age and Months active shows a rising trend, meaning that older employees tend to stay longer with the company as can be confirmed by the correlation value of 0.67 on the heatmap where we found out as age increases months active also increases.

To also give an example of the strength of the relationships from the subplot comparing distance from work and months active. The dots here are more spread out. This indicates a weak relationship. We can also see in the heatmap its correlation is -0.38 which can explain why the trend of the dots are sloping downward and is more spread out.

Assignment 2

Multicollinearity

Multicollinearity refers to the problems that arise when some of the independent variables correlate very highly with each other. This can lead to problems in the estimation of coefficients. Often this makes them unreliable, and the right conclusion can not be drawn.

To check for multicollinearity the following steps were taken:

1. The variance Inflation factor (VIF) was used o check for multicollinearity within the independent variables. The VIF measures how much a variable is influenced by the other variables. If a VIF score is higher than five (5), it means there might be multicollinearity issues.
   From the results 3 variables had the highest VIF score:
   - Height (28.8)
   - Weight (157.99)
   - BMI (147.15)

   The solution was to remove both Height and weight since they can be both inferred from the BMI.

2. The next step was recalculating the VIF but without the height and weight. The results all had a VIF score of less than 5 which meant multicollinearity was successfully handled.

APA Table

Regression Results: Dependent Variable – Months Active

VARIABLE	(1)	(2)
ABSENT_HOURS	-0.002	-0.006
	(0.007)	(0.008)
AGE	0.267***	0.276***
	(0.019)	(0.020)
BMI	-1.109***	0.207***
	(0.272)	(0.030)
CHILDREN	-0.106	-0.066
	(0.101)	(0.104)
DISCIPLINED_YES	-1.152***	-1.189***
	(0.425)	(0.437)
DISTANCE_FROM_WORK	-0.603***	-0.504***
	(0.093)	(0.092)
HEIGHT	-0.489***	
	(0.087)	
INTERCEPT	82.811***	-1.315
	(14.903)	(0.898)
PETS	-0.951***	-0.903***
	(0.082)	(0.084)
SOCIAL_DRINKER_YES	2.281***	1.958***
	(0.231)	(0.229)
SOCIAL_SMOKER_YES	2.728***	2.338***
	(0.414)	(0.420)
WEIGHT	0.451***	
	(0.095)	
OBSERVATIONS	666	666
R^2	0.677	0.655
ADJUSTED R^2	0.671	0.651

RESIDUAL STD. ERROR	2.509 (df=654)	2.586 (df=656)
F STATISTIC	124.370*** (df=11; 654)	138.618*** (df=9; 656)

Note: *p<0.1; **p<0.05; ***p<0.01

Assignment 3 .1

Independent variables

In the model, the independent variables are the pieces of information that influence how long an employee will stay at the company before quitting. For the model to be able to accurately make predictions it is necessary to feed it with the factors that have an influence on what you want to predict. These factors are called independent variables. Below are all the independent variables used and how they may be influencing the number of months active.

Distance from work

Living far from work can be difficult causing employees who commute further to have lesser months active due to the stress.

Age:

Older employees may prefer job stability and stay longer; younger ones might be more open to switching jobs.

Disciplined

Being disciplined for bad behaviour might signal issues at work, which could lead to quitting the job earlier.

Children

Employees with more children may prefer stability or may leave if the job doesn't support work-life balance.

Social drinker

Socializing with colleagues might increase a sense of belonging, leading more months active.

Social smoker

Just like drinking, social smoking could reflect workplace bonding, which might influence retention.

Pets

Having pets might affect work-life balance needs or flexibility preferences, impacting how long someone stays.

BMI

This might be linked to general well-being, which could influence attendance or job satisfaction.

Absent hours

More frequent absences might suggest dissatisfaction which could lead to quitting sooner.

Assignment 3.2

To figure out which group is right, we look at the standardized coefficients from the regression model. Since all the variables are scaled the same way, this lets us fairly compare how strongly each one affects how long employees stay at the company.

After running the model and analysing the results, we looked at three factors that management believed might have the biggest influence on how long employees stay at the company:

Age (+1.79)

- For every standard increase in age, employees stay around 1.79 months longer.
- This is a large and strong positive effect indicating older employees tend to stay longer.

Social Drinker (+1.96)

- Employees who go out for drinks with colleagues stay around 1.96 months longer than non-drinkers.
- This is also a strong and positive effect.

Distance from Work (-0.68)

- Meaning: Employees who live further away tend to stay about 0.68 months less.
- The effect is real but smaller than Age and Social Drinker.

While Social Drinker has the highest coefficient, Age has a bigger long-term impact because it can vary more widely across people, giving it a stronger and more consistent influence overall. Therefore, the group that believed Age has the strongest effect is correct.

Assignment 4.1

Linear Regression

- cross validation with 5-fold
- mean absolute deviation.

the average prediction error is 2.0679.

Assignment 4.2

Best alpha = 12

ridge regression

- Normalized data
- Grid search
- cross validation with 5 folds
- best alpha

The average prediction error with ridge regression is **2.0434.**

Assignment 4.3

Neural Network

The average prediction error for the neural network with the provided hyperparameters is 0.4889

Evaluation of the results with cross validation with 5 folds and mean absolute error scoring.

Model	Average Prediction Error
Linear Regression	2.0679
Ridge Regression	2.0401
Neural Network	0.4889

The neural network performed significantly better than both linear regression and ridge regression models. While the regression models were off by about 2 months on average, the neural network made predictions with less than half a month of error.