ISMM Module Assessment Cover Sheet

Module	M3: Data Modelling
Supervisor/Assessor	
Candidate number	NA11
Date submitted	

Feedback

Final Grading for student

	Fail		**Marginal Fail**	**Pass**			**Distinction**	
	Unacceptable (0-49)	Significant weaknesses (50-54)	Weaknesses (55-59)	Satisfactory (60-64)	Good (65-69)	Very good (70-74)	Excellent (75-79)	Outstanding (80-100)
Indicative grade	F	E	D	C	B	B+	A	A+

Penalty for lateness	20% of marks per week or part week that the work is late

3. Statistical model to predict machine defects

Summary of approach

This task seeks to determine to 'most appropriate' statistical model to predict the number of defects in any given machine using the dataset provided.

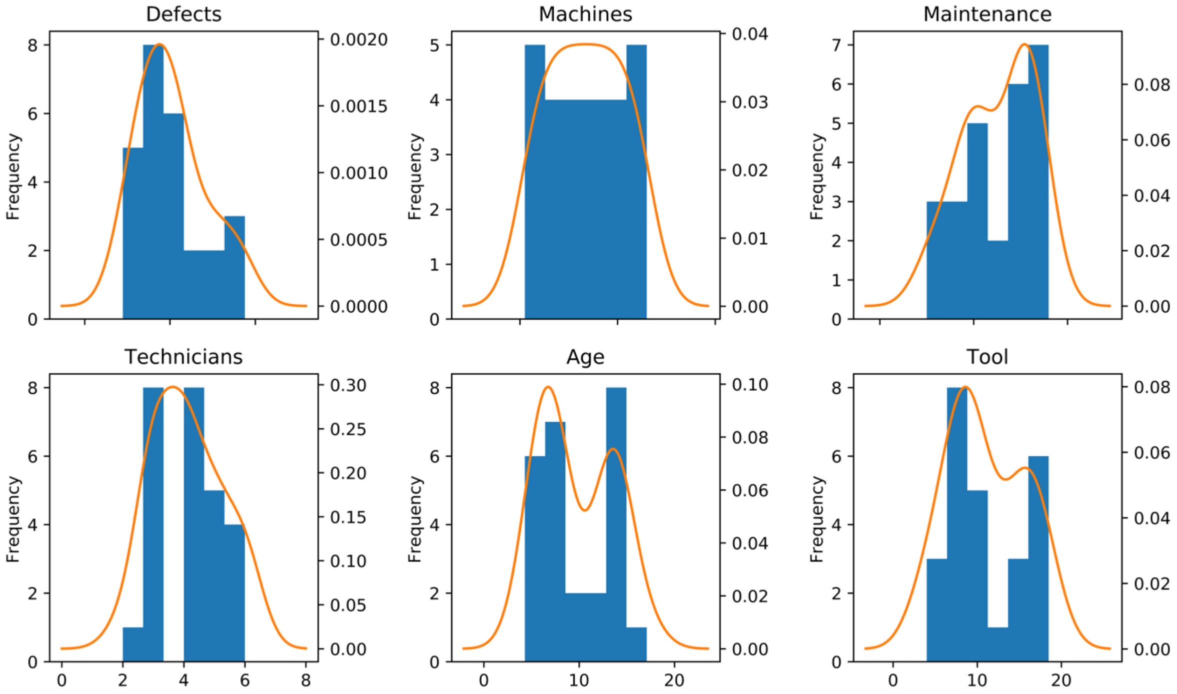
The model development followed an eight-step process and includes an addendum of additional statistical metrics that, while interesting, are not necessary to understand the model development.

1. Gain a first impression of the statistical distributions of each variable in the data set
2. Initial correlation analysis of each variable in the data set with 'Defects'
3. Apply transformations on potential non-linear correlations identified in Step 2.
4. Detailed correlation analysis of each variable and transformation with 'Defects'
5. Create (multiple) linear regression models to calculate following a bottom-up method prioritised by each feature's correlation coefficient
6. Output and compare statistical performance metrics of each computed model
7. Choose a final model and evaluate its performance
8. Specification of final model

A. [Addendum] Compute more detailed statistical performance metrics and plots for each model to ensure validity

The highlights of these steps are discussed below. The full Python Jupyter notebook included in the submission folder follows the same structure and can be referenced for more detail.

1. First impressions of statistical distributions of each variable



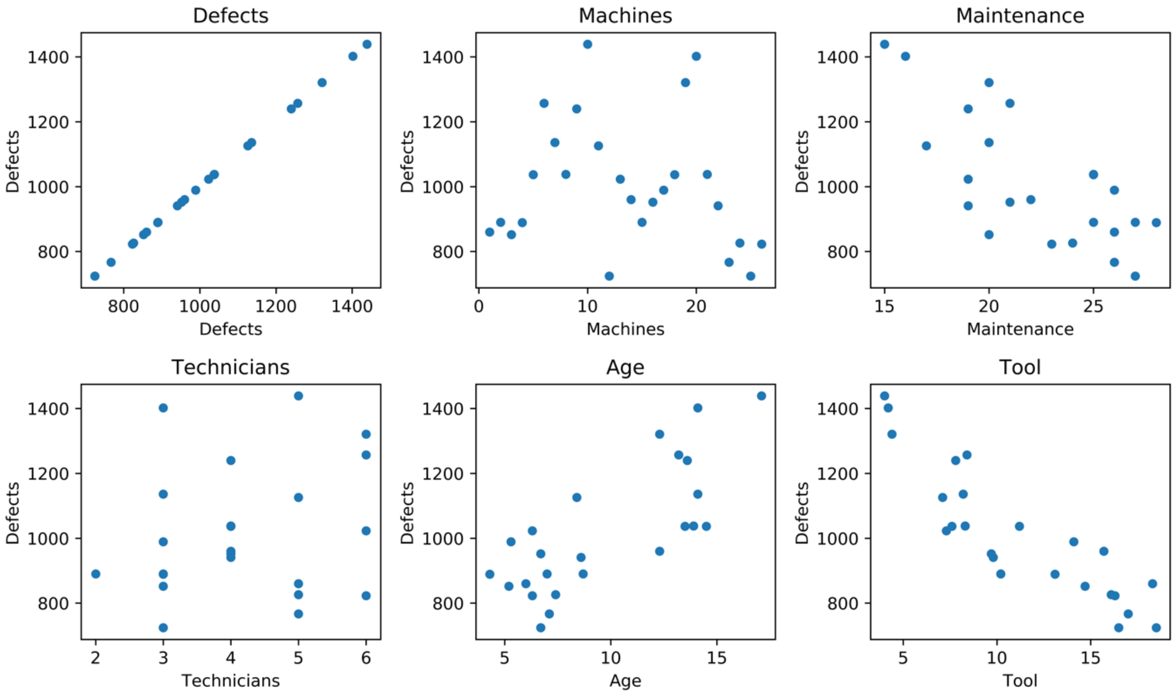
Conclusions from feature distributions

The above distribution plots include both a histogram and Gaussian Kernel Density Estimate (KDE) plot for each feature (variable) in the dataset. A KDE (as the name implies) estimates the variable's probability density function—the continuous equivalent to a the inherently discrete histogram.

Due to the limited sample size (26) the KDEs have very low peak densities (for context, the area under the KDE curve is 1), this means that the KDEs do *not* provide particularly reliable pictures of the population's distribution.

Discussion of each histogram
(reading from left-to-right, then top-to-bottom)
1. **Defects:** Unimodal distribution, right skew due to an higher number of machines in the highest defect bin than would be expected of a normal/Gaussian distribution.
2. **Machines:** Machine number is a unique identifier for each machine, therefore the histogram of machine number frequency contains *no useful information* since each identifier is unique and has no inherent meaning.
3. **Maintenance:** Potentially bimodal distribution, but centre dip could be a result of binning. Too few data points to be sure.
4. **Technicians:** Again, potentially bimodal distribution, but centre zero could be a result of binning. Too few data points to be sure (may be pseudo-normal).
5. **Age:** Likely bimodal distribution, meaning there are a group of 'new' machines and a group of 'older' machines.
6. **Tool:** Again, likely bimodal distribution. This means they can be grouped into one a group of machines undergoing frequent tool changes, and another group subjected to fewer tool changes.

2. Initial correlation analysis



Conclusions from initial scatter plots
The above scatter plots show that 'Defects' shows some significant correlation with the 'Maintenance', 'Age' and 'Tool' variables. Multiple linear regression is suitable method to build a statistical model to predict machine defects in this case, as it can be used to describe a multi-variate relationship with relative ease and effectiveness. If any individual relationships are non-linear, they can be transformed ahead of the linear regression calculation.

Discussion of each scatter plot

(reading from left-to-right, then top-to-bottom)

1. **Defects:** Ignore, plot is simply a result of convenience for the loop.
2. **Machines:** The machine number is ***not correlated*** with the number of yearly defects. This is expected as machine number is an effectively random assignment and has no meaning other than being a unique identifier.
3. **Maintenance:** The number of maintenance operations performed is ***somewhat negatively correlated*** with the number of yearly defects. This makes sense as one would expect more frequently maintained machines to produce fewer defective parts. The scatter is quite loose, but the relationship appears to be slightly non-linear.
4. **Technicians:** The number of technicians responsible for a machine appears to have ***no clear correlation*** with the number of yearly defects. It is possible on a larger sample set may show a stronger correlation, but this sample of 26 does not.
5. **Age:** The age of a machine is ***somewhat positively correlated*** with the number of yearly defects. This again seems reasonable as one would expect older machines to perform more poorly.
6. **Tool:** The average number of tool changes per month performed by a machine is quite ***strongly negatively correlated*** with the number of yearly defects (Defects). This suggests that the longer a tool is in use, the more frequently it produces defective parts. This relationship appears to be non-linear.

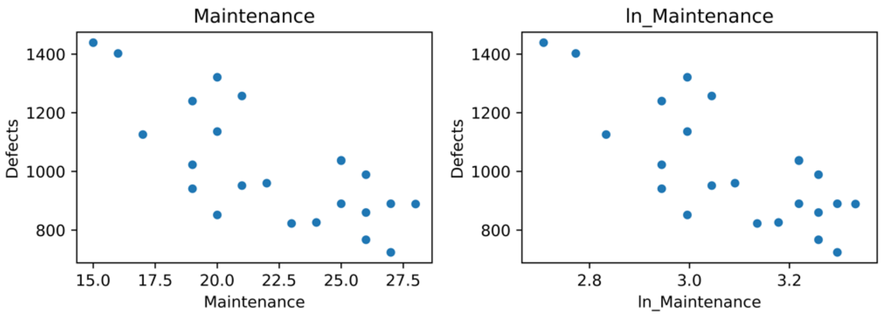
3. Apply transformations to potentially non-linear correlations

As mentioned above, the scatter plots for 'Maintenance' and 'Tool' appear to perhaps be non-linear. Both look like they could be better fitted with a straight line if the x-axis was on a logarithmic scale, this results in the transformation:

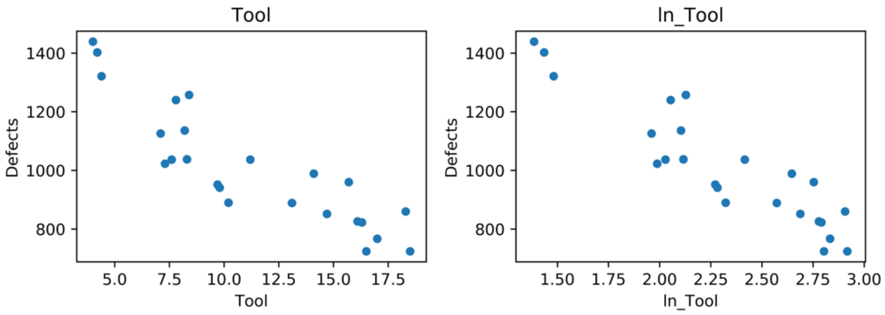
$$y=a+ln(x)$$

where:

- y is 'Defects'
- x is the feature (i.e. 'Maintenance' or 'Tool'
- $ln()$ is the natural logarithmic function



The scatter plots above show that log transformation does not appear to have a significant impact on the linearity of the relationship between 'Maintenance' and 'Defects'.

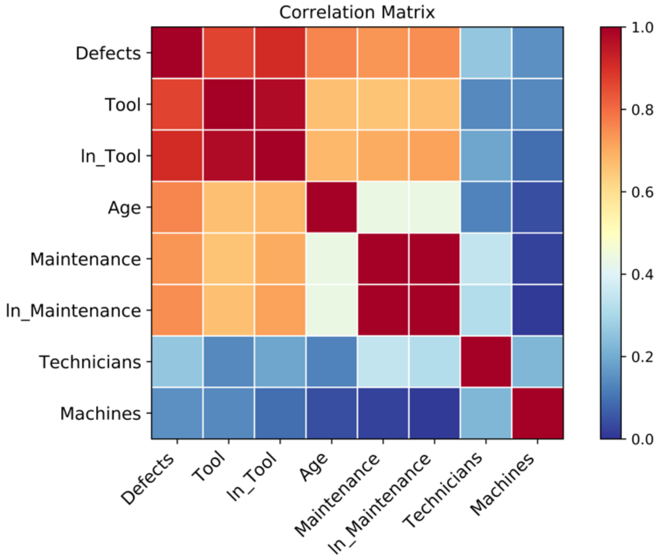


The scatter plots above show that log transformation appears to result in a significantly more linear scatter plot of the relationship between 'Tool' and 'Defects'. This transformation is somewhat justified by domain knowledge/‘common sense’, given that one would expect there to be a lower asymptotic limit on the number of defect as tool changes continue to increase.

4. Detailed correlation analysis

	Defects	*Tool*	*ln_Tool*	*Age*	*Maint.*	*ln_Maint.*	*Techn.*	*Machines*
Defects	1							
Tool	-0.867	1						
ln_Tool	-0.908	0.975	1					
Age	0.760	-0.661	-0.676	1				
Maintenance	-0.735	0.655	0.702	-0.434	1			
ln_Maintenance	-0.747	0.662	0.717	-0.436	0.997	1		
Technicians	0.259	-0.138	-0.189	0.128	-0.345	-0.319	1	
Machines	-0.151	0.138	0.094	-0.042	-0.021	-0.011	0.227	1

The above table of all correlation coefficients can be more easily visualised with a coloured correlation matrix grid:



Conclusions drawn from correlation coefficients
Potential multicollinearity

With correlation coefficients of 0.65-0.7, it is clear that 'Age' and 'Maintenance' (and it's transform) are both somewhat correlated with 'Tool' (and its transform). It is important to be mindful that these feature variables may not be entirely independent–the T-stat value of the features in the final model need to be statistically significant.

Of course, a feature and its transform cannot be included in the same model as they are highly colinear and directly dependent on each other.

Features most fit for inclusion in the (multiple) linear regression model

Feature	*Correlation to Defects & best transformation*
Tool	Linear correlation good (-0.87) Log transform substantially better (-0.91)
Age	Linear correlation good (0.76)
Maintenance	Linear correlation good (-0.74) Log transform not substantially better (-0.75)
Technicians	not strongly correlated (0.26)
Machines	not strongly correlated (-0.15)

The correlation coefficient for the log-transformed values for 'Tool' is substantially better than the linear relationship (from -0.87 to -0.91). Along with this, the scatter plot for the of Defects vs. ln(Tool) appears much closer to a truly linear correlation. This leads to the conclusion that log-transforming the Tool variable is suitable and appropriate for improvement of fit.

Age is reasonably linearly correlated with defects (0.76) and is suitable for use in the model.

Conversely to the 'Tool' case, the log transformation of 'Maintenance' does not substantially change the correlation coefficient or the scatter plot (from -0.74 to -0.75). This agrees with the earlier scatter plots and leads to the conclusion that it is probably best to keep the data in it's original form (i.e. un-transformed) to maintain data integrity and avoid erroneous manipulation.

'Technicians' are not sufficiently correlated with 'Defects' to warrant inclusion in a linear regression model but will be checked as it is a true independent variable.

'Machines' is not a random variable without inherent meaning and should not be included in the model.

5. Create linear regression models following bottom-up method

The models to be compared follow a bottom-up development approach: adding features one at a time in order of decreasing values of correlation coefficient.

Model 0: Features: 'Tool' This model is a simple linear regression model using only 'Tool' to predict 'Defects'. Note that the original (i.e. un-transformed) variable is used in this first model to provide a baseline upon which to build.

Model 1: Features: 'ln(Tool') This model is also a simple linear regression model but now uses the transformed value of 'Tool'—ln('Tool')—to predict 'Defects'. This transformation resulted in a higher correlation coefficient and will result in a better fitting model.

Model 2: Features: 'ln(Tool)', 'Age' This model is a multiple linear regression model building upon the previous. It adds the 'Age' feature, the feature with the next-highest correlation coefficient.

Model 3: Features: 'ln(Tool)', 'Age', 'Maintenance' This model is, again, a multiple linear regression model building upon the previous. It adds the 'Maintenance' feature, the feature with the next-highest correlation coefficient and the last independent feature with a reasonably high correlation coefficient.

Model 4: Features: 'ln(Tool)', 'Age', 'Maintenance', 'Technicians' This final model a multiple linear regression model that includes all independent features in the provided dataset. It adds the 'Technicians' feature, which is not well-correlated with 'Defects'.

6. Output and compare performance of each model

Evaluation of model coefficient t-statistic values

	Model 0	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Intercept	26.7991	21.7833	10.8719	11.8322	9.82072
ln(Tool)		-10.5934	-6.93496	-4.56196	-4.52992
Age			2.56942	2.92029	2.8625
Maintenance				-2.14805	-1.83213
Technicians					0.607604
Tool	-8.50513				

All features in model 0-3 have $|t\text{-}stat|>2$. This means that all features are statistically significant in models 0-3 and the models are at least robust on these grounds.

In model 4, 'Maintenance' and 'Technicians' have $|t\text{-}stat|<2$ (orange cells). This means that these features fail to reject the null hypothesis positing that their contribution to the model is not statistically significant. As a result, this model is not valid due to its inclusion of statistically insignificant features.

Comparison of model coefficients

	Model 0	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Intercept	1428	1927.9	1607.8	1693	1642.5
	(53.284)	(88.505)	(147.89)	(143.08)	(167.25)
ln(Tool)		-398.39	-318.61	-247.07	-249.7
		(37.607)	(45.943)	(54.159)	(55.123)
Age			13.931	14.763	14.687
			(5.422)	(5.0553)	(5.1309)
Maintenance				-11.437	-10.389
				(5.3242)	(5.6706)
Technicians					8.1627
					(13.434)
Tool	-38.141				
	(4.4844)				

Key:

- number – Coefficient
- (number) – Standard error of coefficient

Comparison of model performance metrics

	Model 0	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Adjusted R2	0.740	0.816	0.851	0.871	0.868
p(F-stat)	1.05e-08	1.58e-10	1.17e-10	1.43e-10	9.96e-10
RMSE Residuals	100.6	84.6	76.1	70.8	71.8

This table provides three key performance metrics for the models:

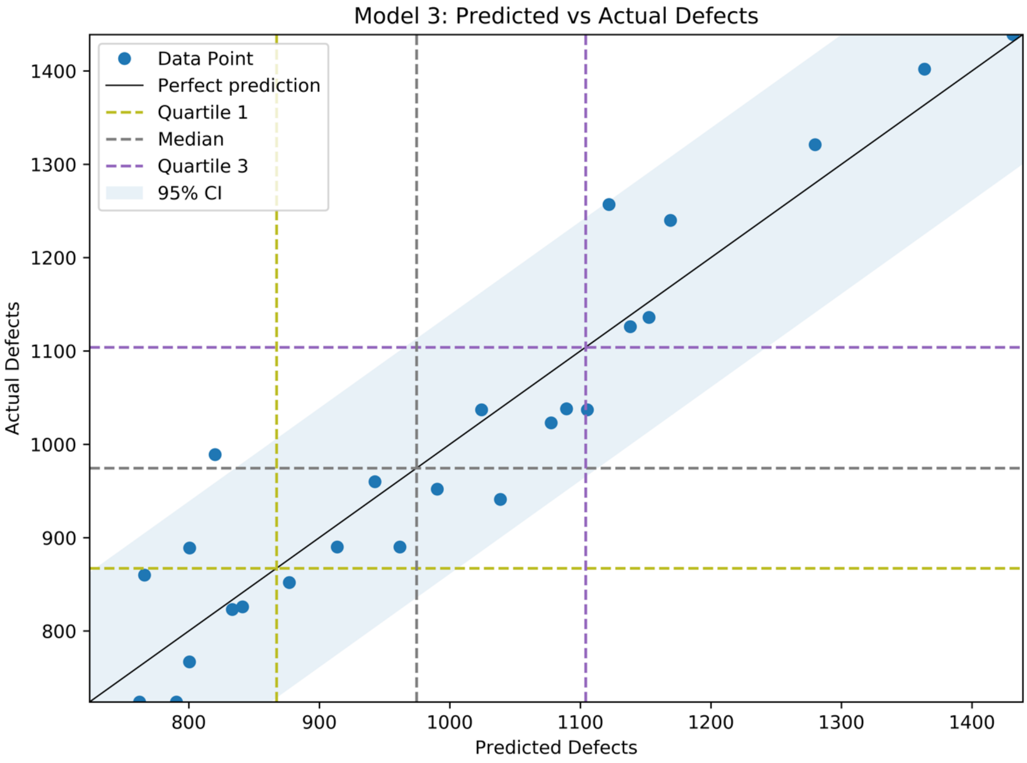
- Adjusted R2 indicates how much variance is accounted for by the model.
- p(F-stat) is the probability of the model's F-statistic is the probability that the model fails to reject the null hypothesis (i.e. that the model is not statistically significant).
- RMSE residuals is the root mean squared error of the residual error in the model (also called standard error).

The adjusted R2 and RMSE residual values show that each model improves on the previous, except for model 4. This is interesting because it means the inclusion of 'Technicians' in model 4 actually resulted in a worse performing model (without even considering the fact that the model is statistically insignificant due to two of its features failing the t-stat test.

The p(F-stat) values show that all models confidently reject the null hypothesis. The model with the highest value still only having a 0.000001% probability of being statistically insignificant (model 0).

7. Choose final model and evaluate its performance

Model 3 is chosen as the final model as it has the highest adjusted R2 value while ensuring all the features pass the t-stat test and are statistically significant. The number of features included in the model tangibly reduce the RMSE of residuals, allowing for the tightest prediction confidence intervals.



The plot above shows that the model's achieved standard error (RMSE of residuals) results in a confidence interval that means the model predictions usefully predict the number of defects a machine will produce in a year *roughly* accurate to the quartile into which the number of defects will fall.

The only datapoints that are predicted in the incorrect quartile fall near the quartile boundary, which is to be expected as the confidence interval spans both sides of the prediction (the closer the prediction is to a boundary, the more likely the confidence interval will run over the boundary).

8. Specification of final model

The final model specification is provided in the table below:

	coeff.	*std err*	*t*	*[0.025]*	*[0.975]*
Intercept	1693.00	143.08	11.83	1396.26	1989.74
ln(Tool)	-247.07	54.16	-4.56	-359.39	-134.75
Age	14.76	5.06	2.92	4.28	25.25
Maintenance	-11.44	5.32	-2.15	-22.48	-0.40

The table can be interpreted as follows:

The final multiple linear regression model developed to predict defects is:

$$Defects = 1693 - 247.07 \cdot \ln(Tool) + 14.76 \cdot Age - 11.44 \cdot Maintenance$$

The upper and lower 95% confidence intervals for the prediction are:

$$Defects_{lower\ bound} = 1396.26 - 359.39 \cdot \ln(Tool) + 4.28 \cdot Age - 22.48 \cdot Maintenance$$
$$Defects_{upper\ bound} = 1989.74 - 134.75 \cdot \ln(Tool) + 25.25 \cdot Age - 0.4 \cdot Maintenance$$

A. Addendum: More detailed statistical performance metrics/plots

Regression plots for each feature in each model

The primary conclusion from this section is that the scatter of all residual plots for all features in all models appears random, validating the use of all features in the models.

Influence plots for each model

The primary conclusion from this section is that no single point has undue leverage on the parameterisation of the model: i.e. outliers (if any) are not skewing the model unreasonably.

Leverage plots for each model

These plots show that data points 5 and 16 (i.e. machines with ID no. 5 and 16) are outliers in the model, but they do not have disproportionate leverage on the model's parameters.

It is also worth noting that model 1 is highly influenced by data points 9, 18 and 19 (i.e. machines with ID no. 9, 18 and 19). This shows disproportionate reliance on a few data points and challenges the validity and robustness of model 1. However, model 1 was not chosen as the final model so this is not an issue.