Assessment 2a: Questions and Template

Version 1.0
EGH404: Research in Engineering Practice
25 August 2019

About this document

- This document sets out the five (5) questions you are to complete for assessment 2a.
- It also provides the template for your responses.

Your tasks for Assessment 2a

1. Read the questions in this document
2. Edit your responses into this document
   a. Replace the text in *blue italics* with your responses
   b. The associated data and/or original graphs are provided separately in an Excel spreadsheet where necessary.
   c. All necessary calculations and charting should be completed in Matlab.
3. You will submit your responses to this assessment as a PDF or word document which you will upload to TurnItIn. Your document should include:
   a. This template with responses for each question included in the corresponding section.
   b. Your original source code (ie. Matlab scripts) which you generated to complete relevant questions, included either after each question or at the end of the document as an appendix (copy and paste the code into the document, or if using live scripts merge the document PDF with the exported live script PDF). Please note your Matlab code will not be run except under exceptional circumstances but may be used to evaluate the process that you used to answer the question.

Question 1: Recommending a bolt supplier

Your firm works in construction and buys ASTM A325M8S bolts in very large numbers.

You are interested in purchasing bolts from different suppliers and have obtained and tested samples of 200 bolts from three suppliers: Allnutt, Boltzman, Coachers.

You have tested these in your new bolt testing machine, purchased after you had discovered your old bolt testing machine was performing erratically. (For the purposes of this question, you are safe to assume that the new machine works perfectly and that its results are in MPa.)

Your firm wants to ensure that the tensile strengths of the bolts it buys are

- as high as possible
- as consistent as possible (i.e., the variability of bolt strengths is as *low* as possible)

It is critical that the tensile strength of these bolts exceeds 830MPa.

You should draw on the unit content concerning summary statistics to answer this question.

On the basis of the measurements recorded in Bolts.csv, which supplier (if any) would you recommend, and why? (Provide any code or visualisations you use to justify your response.)

The following text will demonstrate the methodology in choosing a manufacturer to supply the company with the ASTM A325M8S Heavy Hex Bolts. Factors such as; certifying that the tensile strengths are of the highest rating whilst ensuring that these results are not bolstered by outliers and assessing the variability of bolt strengths. The assessment will use the measurements provided in *Bolts.csv*.

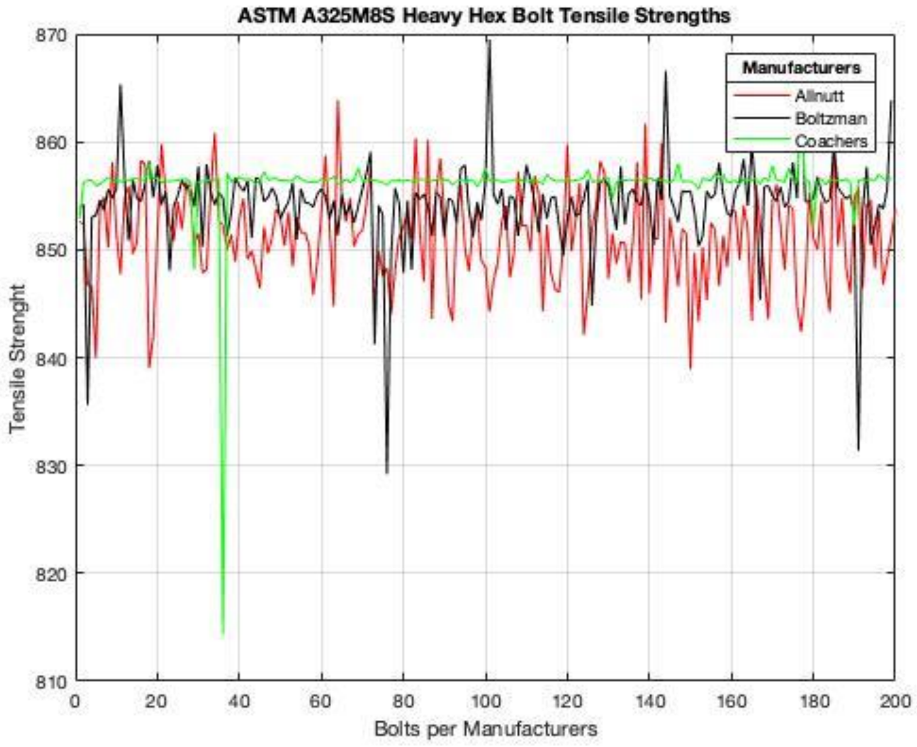


Figure 1: *Raw Data – Bolts.csv*

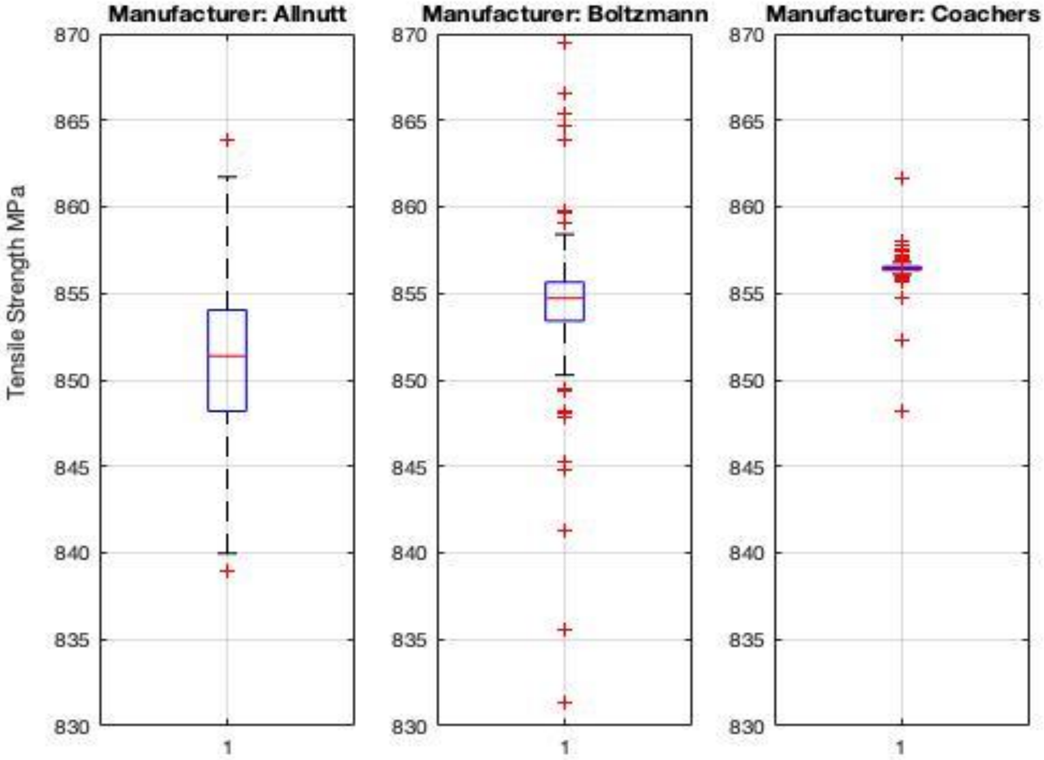


Figure 2: *Box Plot Data showing the Ranges of Tensile Strengths*

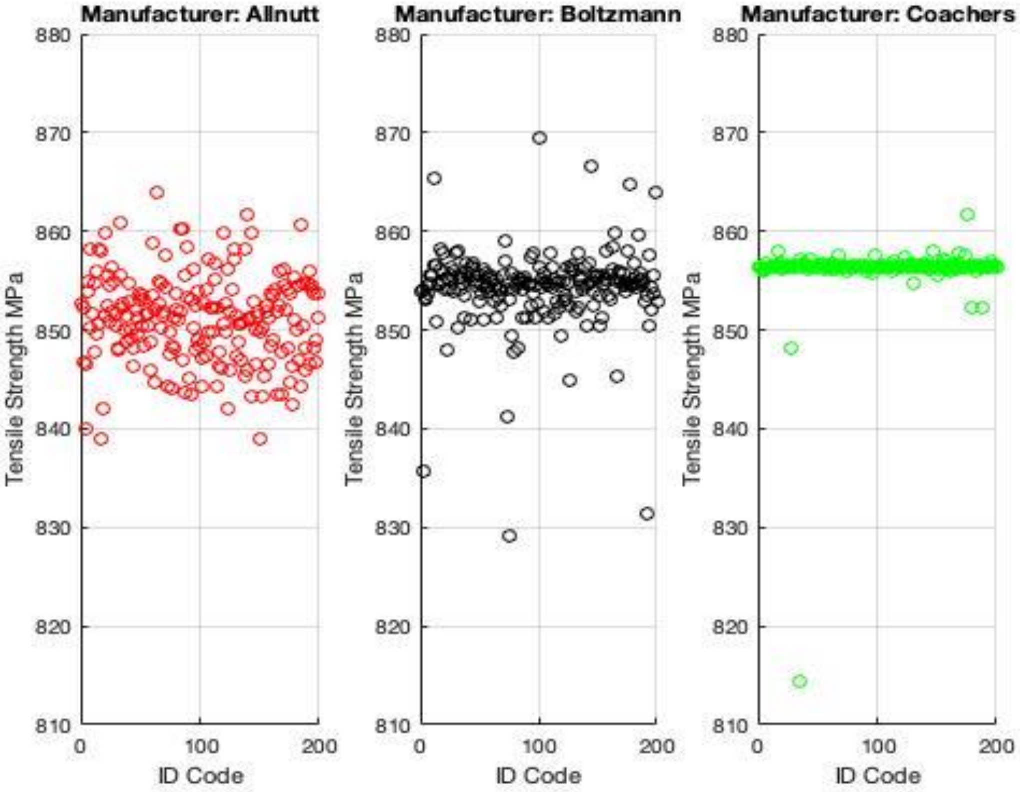


Figure 3: *Scatter Plot showing Tensile Strengths*

To select the manufacturer that suits the criteria a number of conditions were assessed, these being; the minimum and maximum tensile strengths, mean, median and the average absolute deviation (represented in table 1).

Table 1: *Tensile Strength Conditions for Respective Manufacturers*

Tensile Strengths \\ Manufacturers	**Allnutt**	**Boltzman**	**Coachers**
Minimum	838.9322	829.1931	814.3485
Maximum	863.8724	869.4711	861.6270
Mean	851.1000	854.3000	856.2000
Median	851.3811	854.7285	856.4237
Median Absolute Deviation	3.518300	2.229900	0.620600

As the criterion states that bolts with a tensile strength lower than 830MPa are not sufficient these bolts were identified. Bolts with the Identification Numbers ID276 (Boltzman) and ID435 (Coachers) were found to have tensile strengths of 829MPa and 814MPa, respective.

The data presented in Table 1 and Figures 1 and 2 demonstrates that the most suitable supplier of the ASTM A325M8S Bolts is Coachers. Testing of the bolts provided by Coachers demonstrated that the variability of tensile strengths across the 200 bolts was considerably lower than the other two manufacturers (*Allnutt: 12.074 & Boltzman: 7.161 x higher variability*). Furthermore, the Coachers' provided bolts tested for higher strengths across the board as opposed to Allnutt and Boltzman (*Allnutt: 0.589% & Boltzman: 0.197% lower average tensile strengths*).

However, as Coachers provided a bolt that was not to standard (less than 830MPa) they cannot be recommended unless further product requirements are given. This, for example, would include; an allowance that for every 200 bolts provided there will be one defective bolt and this bolt would be identified and removed due to safety.

If it is a requirement that all of the 200 bolts provided are to standard, then it is recommended that Allnutt is selected as the manufacturer.

Question 2: Changes in Air Quality Over Time

The state government's Department of the Environment and Heritage Protection are concerned about air pollution levels in Brisbane city. You have been provided with data for two sites, the Brisbane CBD and South Brisbane, for 2010 and 2016 in the following four files:

- brisbanecbd-aq-2010.csv
- brisbanecbd-aq-2016.csv
- southbrisbane-aq-2010.csv
- southbrisbane-aq-2016csv

You have been asked to advise the Government:

- If air pollution is related across the two sites, and if so, has this relationship changed between 2010 and 2016
- Has the level of air pollution at either site changed from 2010 to 2016? Have these changes across the two sites been consistent?

You have been advised to focus on the PM10 data series.

You should draw on the unit content concerning summary statistics and correlation to answer this question.

(Provide any code or visualisations you use to justify your response.)

The writing below will present an evaluation of the air pollution levels in Brisbane City, Brisbane CBD and South Brisbane for the years 2010 and 2016. The data sets will be used to assess whether air pollution is related across the two sites, and if so, has this relationship changed between 2010 and 2016, furthermore, has the level of air pollution at either site changed from 2010 to 2016, and if so, have these changes been consistent.

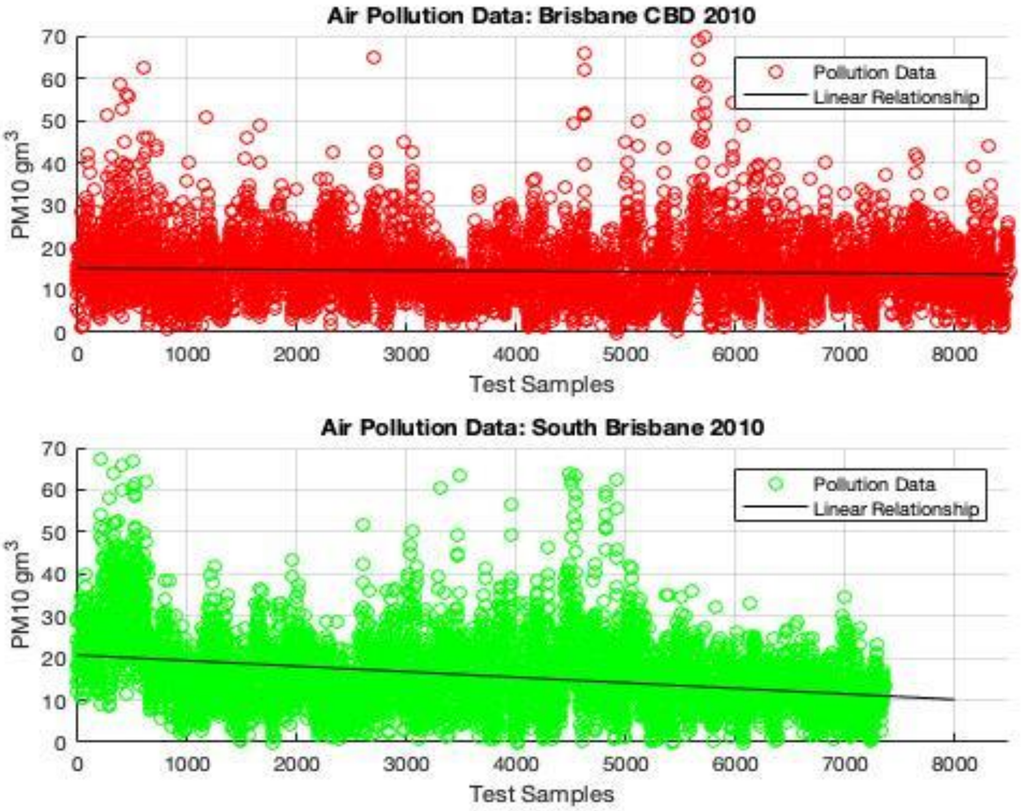


Figure 4: *Scatter Plot showing the Correlation between Brisbane CBD and South Brisbane for 2010*

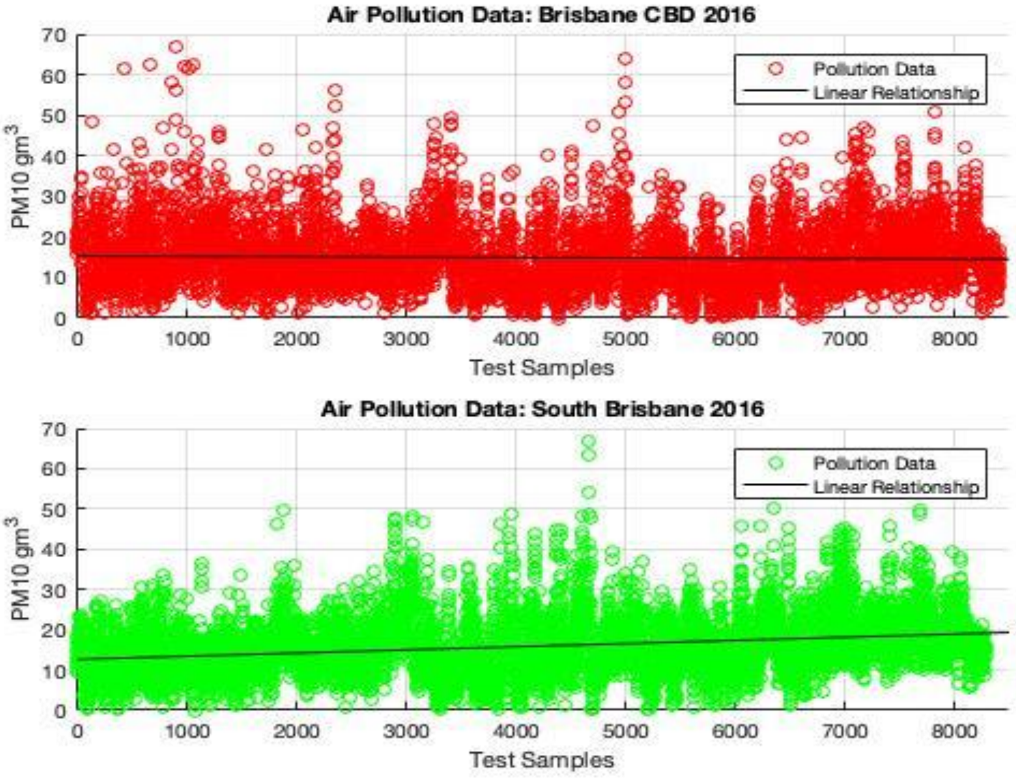


Figure 5: *Scatter Plot showing the Correlation between Brisbane CBD and South Brisbane for 2016 Pollution Levels*

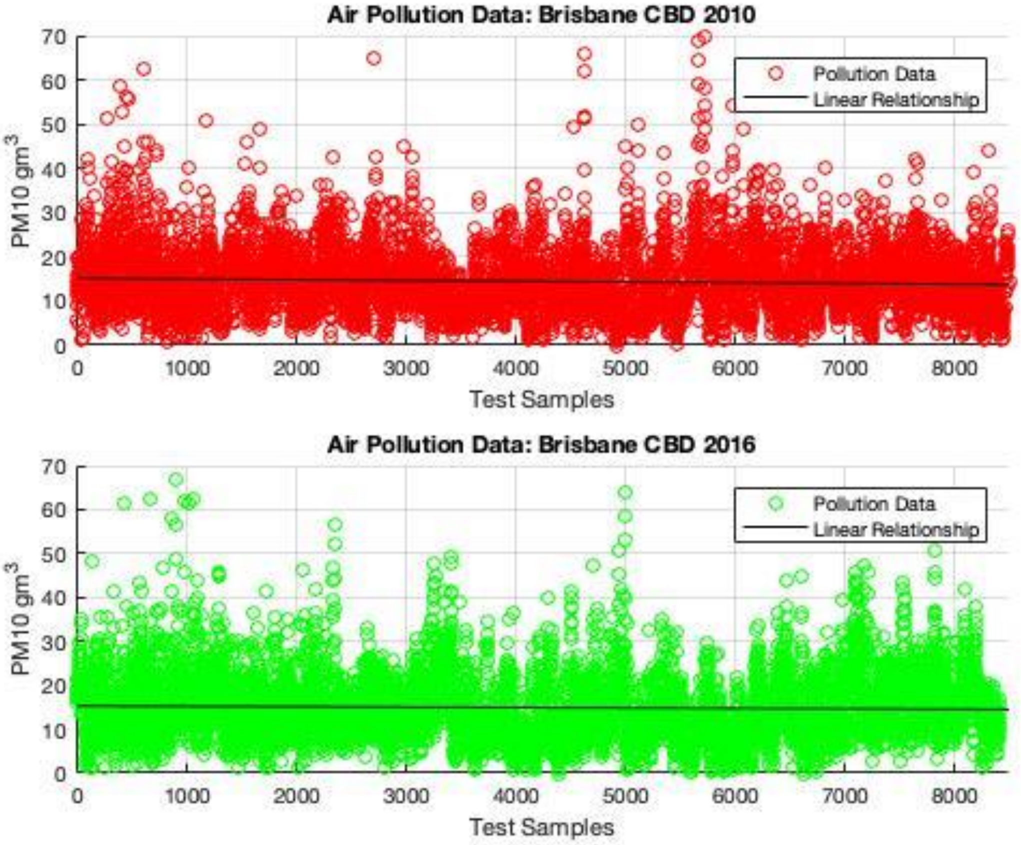


Figure 6: *Scatter Plot showing the Correlation between Brisbane CBD Pollution Levels in 2010 and 2016*

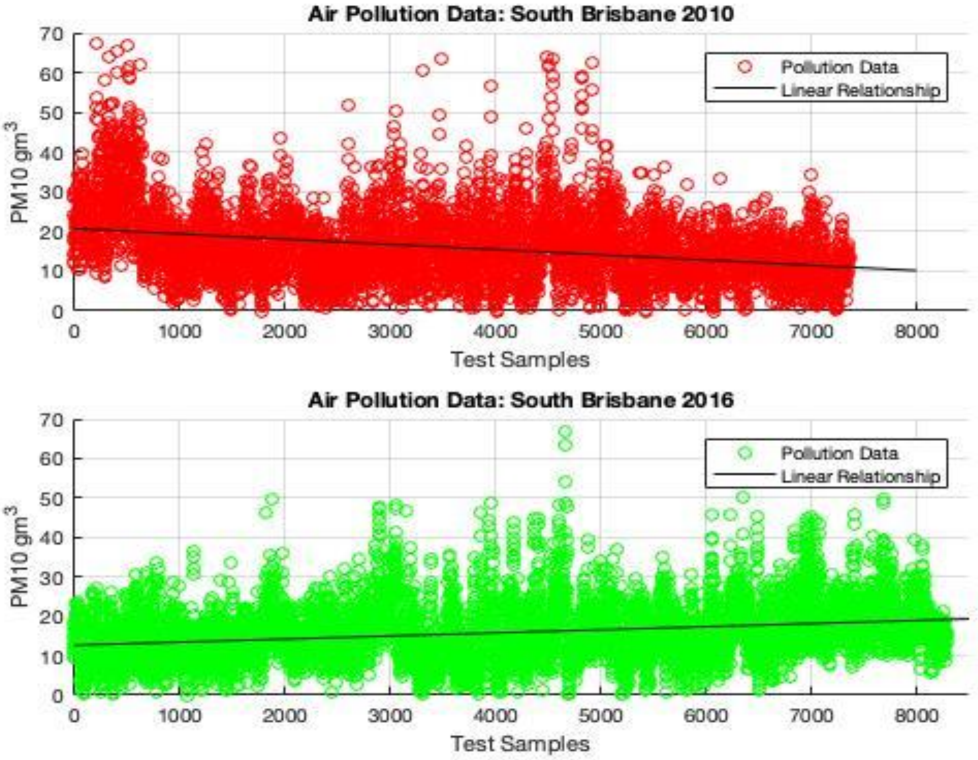


Figure 7: *Scatter Plot showing the Correlation between South Brisbane Pollution Levels in 2010 and 2016*

Table 2: *Shows the Correlation between Brisbane CBD and South Brisbane for 2010 and 2016*

Correlation between Sites for *2010*		
	1.0000	**0.6618**
	0.6618	**1.0000**

Correlation between Sites for *2016*		
	1.0000	**0.6974**
	0.6974	**1.0000**

Table 3: *Demonstrates the Mean, Median and Median Absolute Deviation for Pollution Levels across both Sites for 2010*

Pollution Levels \\ Site	**Brisbane CBD**	**South Brisbane**
Mean	14.43194	15.86508
Median	13.40000	14.50000
Median Absolute Deviation	5.600940	6.704537

Table 4: *Demonstrates the Mean, Median and Median Absolute Deviation for Pollution Levels across both Sites for 2016*

Pollution Levels \\ Site	**Brisbane CBD**	**South Brisbane**
Mean	14.90538	15.89844
Median	13.60000	15.10000
Median Absolute Deviation	5.994581	5.691381

From Table 2, 3, 4 and Figure 4 it can be seen that the correlation in air pollution levels across the sites for 2010 is related. Both sites recorded mean data within 1.43310 points of each other and median data of 0.90000. The variability of the levels in South Brisbane were higher than that in the CBD (by 1.10360).

It can also be seen in Table 4 and Figure 5 that air pollution levels across both sites for 2016 are related. The recorded mean data had a 0.99306 variance between each site and the median data presented a 1.5000 difference. The variability levels in Brisbane CBD recorded 0.30320 higher than that in South Brisbane.

The recorded data for both sites in 2010 can be seen to be slightly related to the data presented in 2016. However, the only abnormality is the lower variability levels in South Brisbane in the 2016 recorded data. Variance in the CBD increased 0.39364 points whereas levels in South Brisbane decreased 1.01316 points. This can also be seen in Figure 7, the linear fit shows to be increasing in the 2016 data, whereas the linear fit is decreasing in the 2010 data.

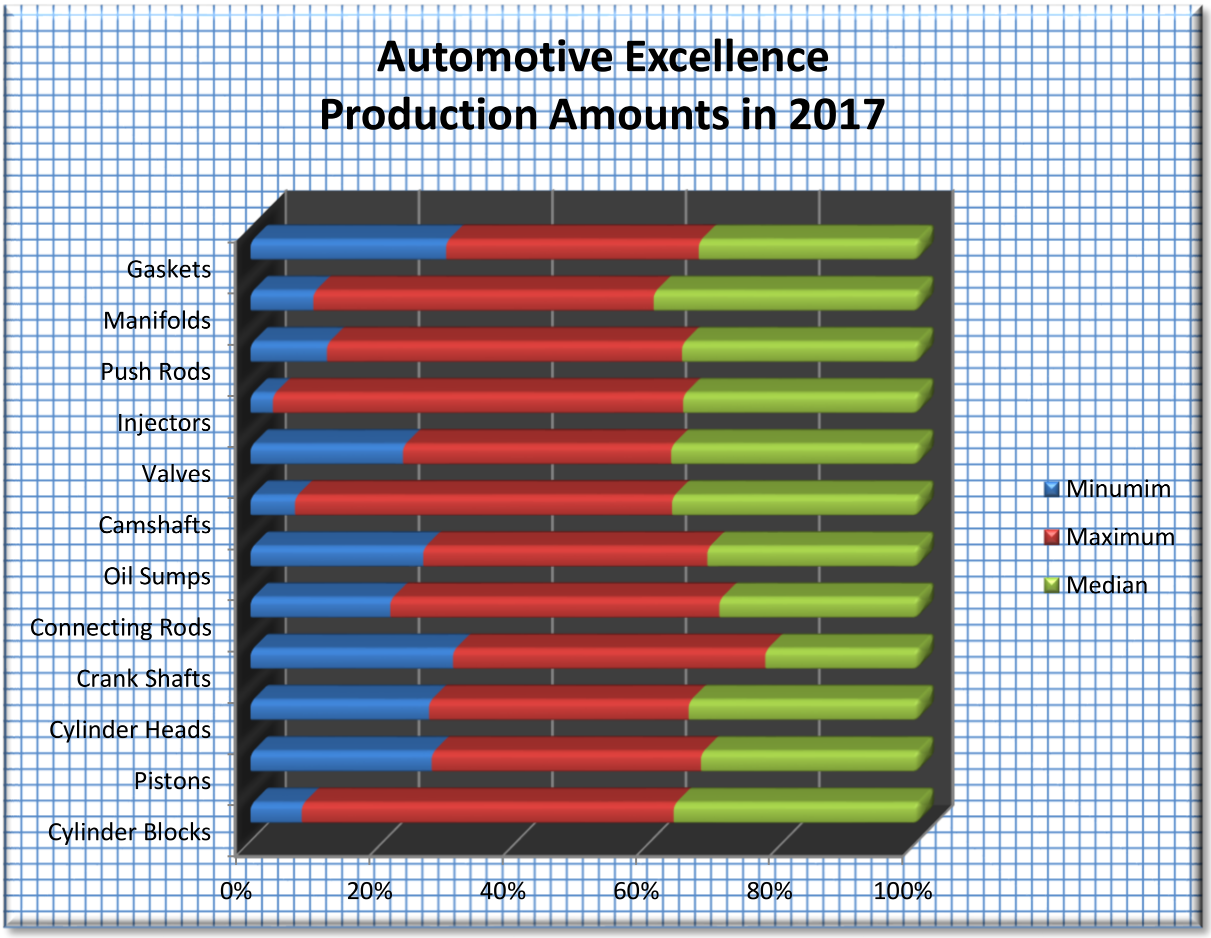
Furthermore, the median data for the Brisbane CBD increased by a slight 0.20000 points and the mean only incremented by 0.473440 points. Whereas the median increased in South Brisbane by 0.60000 points and the mean by 0.03336 points.

These mean, median and median absolute deviation measurements show that the air pollution levels in the Brisbane CBD have remained fairly consistent with a slight increase in variance, 6.78954%, 3.22756% increase in mean and an increase in median of 1.42925%. The relationship for South Brisbane has changed more significantly, an increase of 0.210052% in mean, 4.05405% increase in median and a 16.3466% decrease in variance.

Question 3

Part A

- Automotive Excellence produces a range of components needed to produce automobile engines.
- Production of different component categories varies from month to month.
- The intent of this graph is to summarise the variation in numbers of component categories from month to month in 2017.



A.1 What's wrong with this graphic?

List and briefly explain each of the problems that you detect in the design of this graphic drawing on the principles presented on visualisation and from your wider reading

The list below represents the issues that can be seen in the graph: *Automotive Excellence Production Amounts in 2017*:

- *Issue*: **3D Modelled Graph**
  - The graph paper obscures the y-values and makes the information harder to read
  - The 3D model is unnecessary and makes the data harder to interpretate
  - The y-values do not directly line up with their corresponding data
  - The x-values in percentage make the data difficult to comprehend scale and production amounts required
  - The order of minimum, maximum and median is wrong, this should be in the order: minimum, median and maximum

- ○ The graph doesn’t show the production amounts that vary from month-to-month but only shows the yearly percentage
- ○ The spelling of minimum is incorrect

The graph also introduces too many distractors which decreases the distinctness of the pre-attentive data. That is, it is harder to read the data with the layout of the graph; gridded paper, 3D model, percentage-based x-scale. Furthermore, as the minimum, maximum and median are grouped for their respective product these results are visualised as a group and not individual readings (Spelke, 2004).

A.2 How would you redesign this graphic?

Develop an alternative graphic that addresses the problems you detected

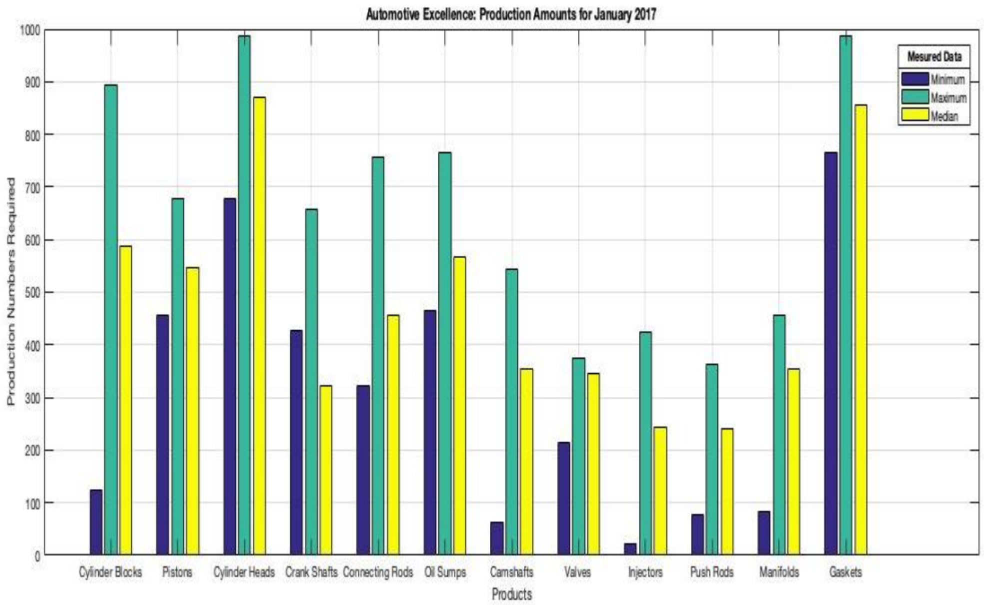


Figure 8: *Re-design of the Graph Presented above*

NB: See appendix A for a larger, clearer image.

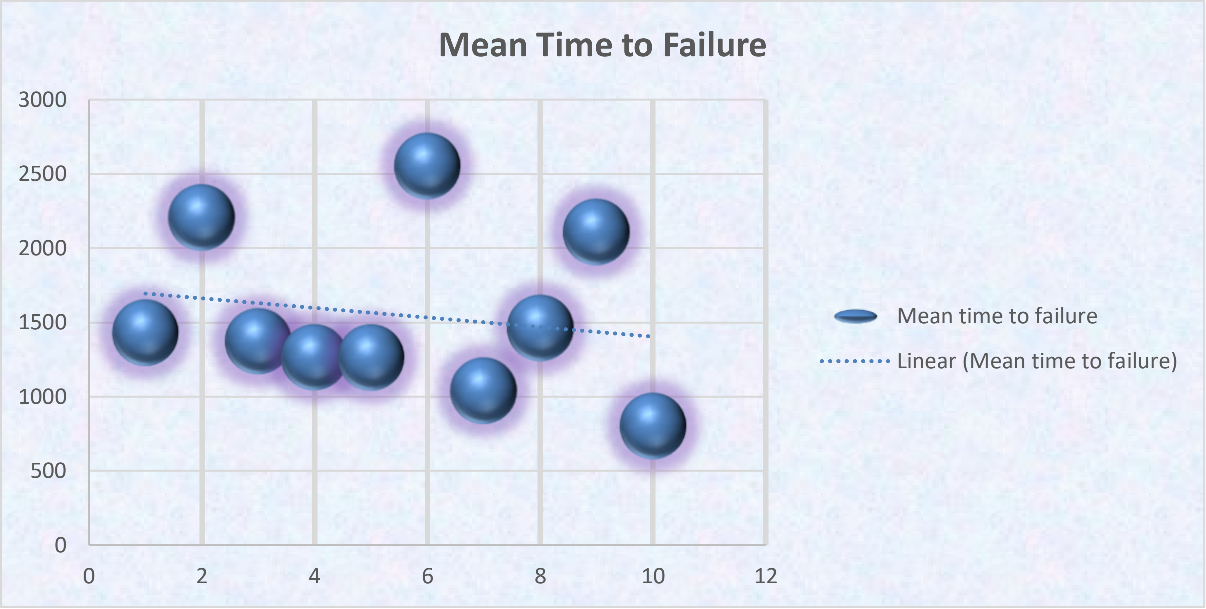
A.3 What do you now observe in your redesigned graphic?

What are the salient problems or features of the data that you now observe?

- The new graph is a bar graph and represents minimum, maximum and median in individual bars
- The y-axis and x-axis represent production numbers and products, respectively
- The order of the data presented should be minimum, median and maximum whereas it is currently, minimum, maximum and median
- The minimum crank shaft data is higher than the median produce, this is incorrect data
- The data collected does not isolate the month-to-month variation and only shows the yearly production numbers
- There is no indication of the required number of components that need to be produced

Part B

- Automotive Excellence has gaskets which come in a range of sizes for different engine blocks and for different components.
- Through a sophisticated product tracking system, Automotive Excellence monitors the working life of these gaskets and have collected time-to-failure data on hundreds of gaskets.
- The Automotive Excellence product quality team have summarised this data by calculating the mean failure times of a sample of each size. The sizes are reported by their overall length in mm.
- The intent of this graph is to help understand whether there is a relationship between gasket size and failure times.



A.1 What's wrong with this graphic?

List and briefly explain each of the problems that you detect in the design of this graphic drawing on the principles presented on visualisation and from your wider reading

The list below represents the issues that can be seen in the graph: *Mean Time to Failure*:

- *Issue*: **Scatter Modelled Graph**
  - The graph does not directly show the gasket sizes in the x-axis so there is no way to determine which 'Mean time to failure' ball represents which size of gasket
  - There is no indication of the measurement of time (y-axis) so the data could represent failure in seconds, minutes, hours, days, etc
  - The size of the markers is too large to read precisely
  - The above dot points suggest that different designs of gaskets are used, these are not indicated in the data, nor the plot
  - As the gasket size is not stated in the x-axis there is no way to determine if there is a relationship between the gasket size and failure times

B.2 How would you redesign this graphic?

Develop an alternative graphic that addresses the problems you detected

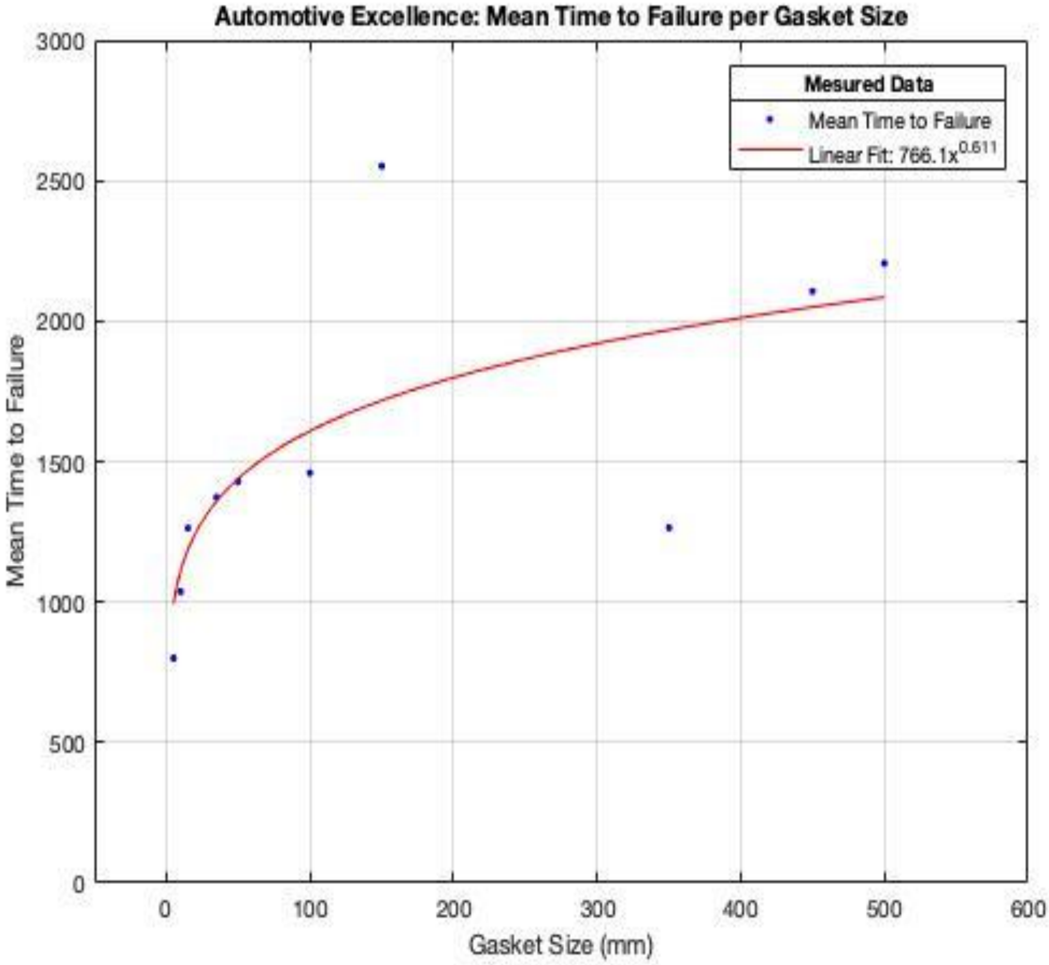


Figure 9: *Re-design of the Graph Presented above*

B.3 What do you now observe in your redesigned graphic?

What are the salient problems or features of the data that you now observe?

- The new graph plots the data in ascending gasket size and a linear fit and logarithmic line is fitted
- The y-axis and x-axis represent the mean time to failure and gasket size, respectively
- The graph could be improved by providing the categories and designs of the gaskets to show the relationships between each individual gasket design and the overall gasket sizes

Question 4: Estimating Concrete Strength

You work for a construction firm who need to be able to accurately predict the compressive strength of concrete given variables including the concrete composition and its age. You are aware that the relationship between these different components and the concrete strength is complex, however you have been asked to investigate how well a simple linear regression model works for prediction. Using the provided data (Concrete_Data.xls), develop models to predict:

- Concrete strength from the single best indicator variable;
- Concrete strength from all variables.

With the second model, determine if any variables are not contributing significantly to the model, and what impact removing these has on prediction performance. Comment on the final model and its accuracy, and whether it would be appropriate to use this model in practice.

You should draw on the unit content concerning correlation and regression to answer this question. Note that you are not expected to use training/validation/testing data splits, although you are welcome to do so. No marks will be lost/gained for using/not using data splits.

(Provide any code or visualisations you use to justify your response.)

The following text will demonstrate how a linear regression model can be used to predict compressive strengths based on certain input variables, such as; concrete composition and its age.

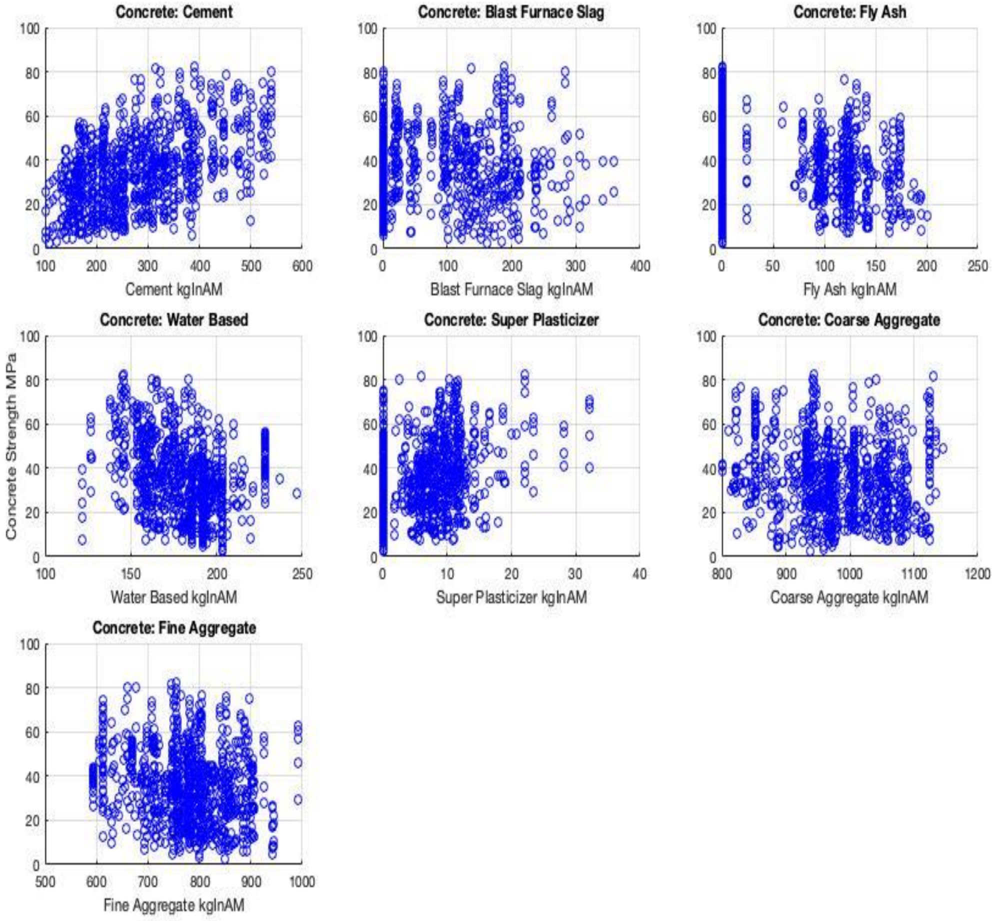


Figure 10: *Scatter Plot Considering all the Concrete Composition Data*

From Figure 10 it can be seen that *cement-based concrete* will be our best indicator as it provides the most consistent data, this will also be supported in the following table.

Table 5: *Correlation between Compressive Strength and the Respective Concrete Component*

Cement		
	1.0000	**0.4978**
	0.4978	**1.0000**
Blast Furnace Slag		
	1.0000	**0.1348**
	0.1348	**1.0000**
Ash Fly		
	1.0000	**-0.1058**
	-0.1058	**1.0000**
Water Based		
	1.0000	**-0.2896**
	-0.2896	**1.0000**
Super Plasticizer		
	1.0000	**0.3661**
	0.3661	**1.0000**
Coarse Aggregate		
	1.0000	**-0.1649**
	-0.1649	**1.0000**
Fine Aggregate		
	1.0000	**-0.1672**
	-0.1672	**1.0000**

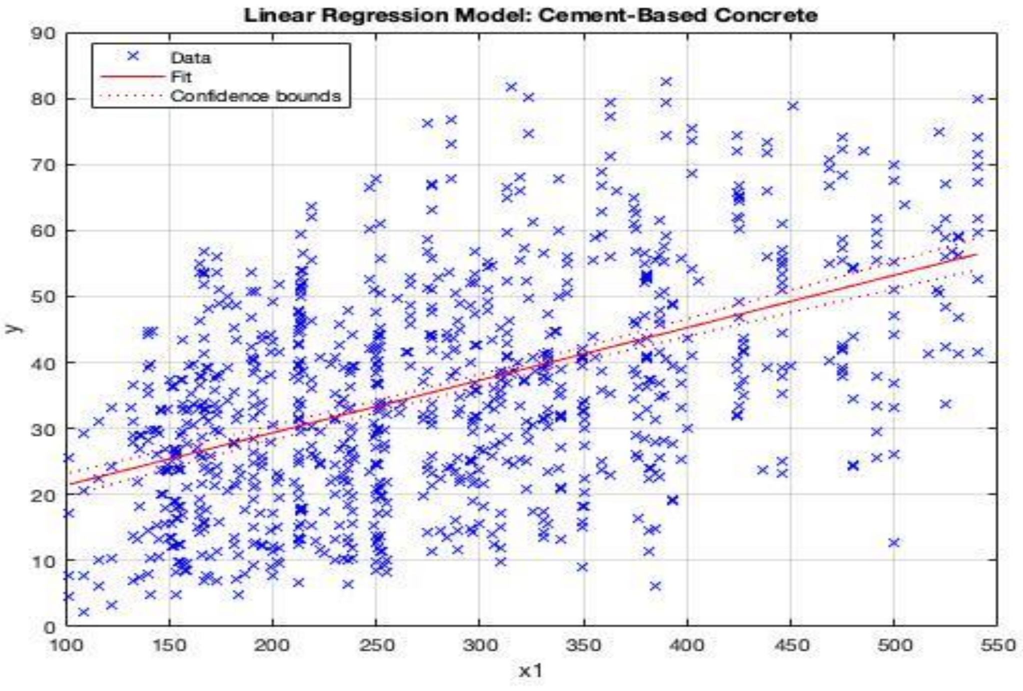


Figure 11: *Linear Regression Model for Cement-Based Concrete*

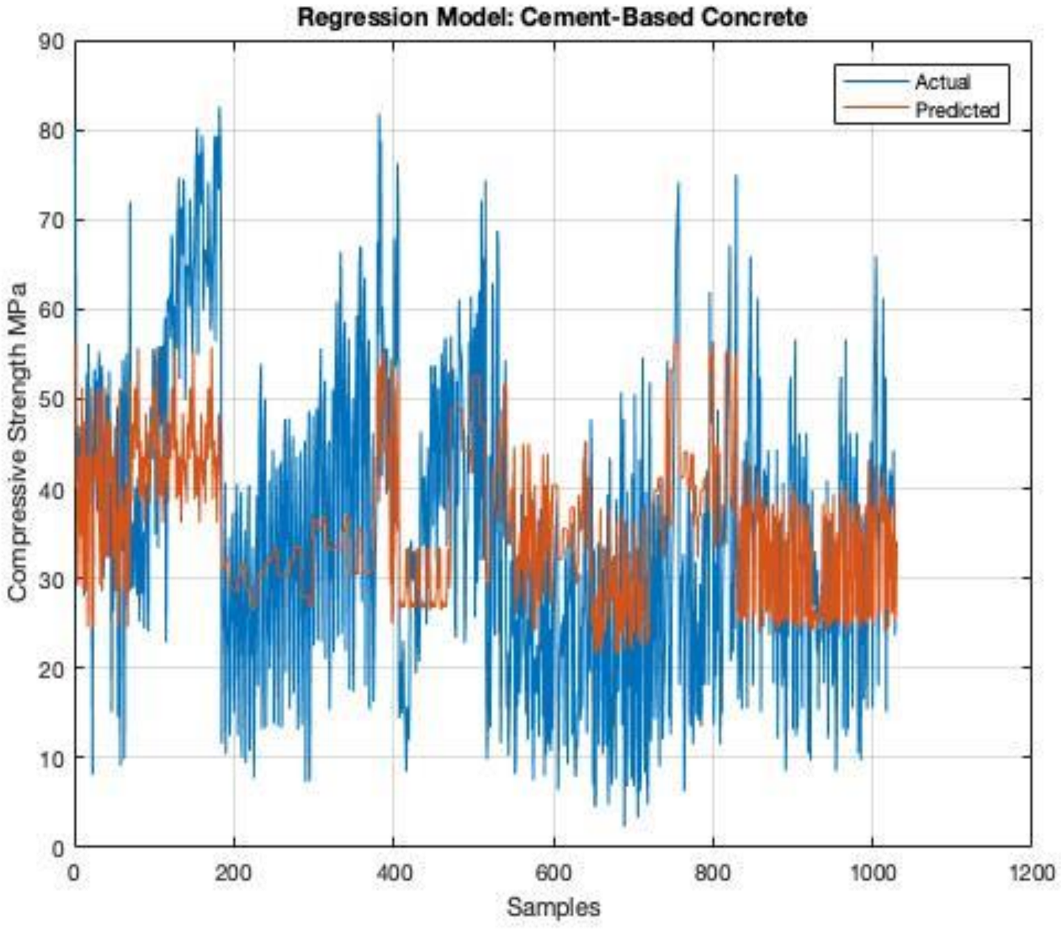
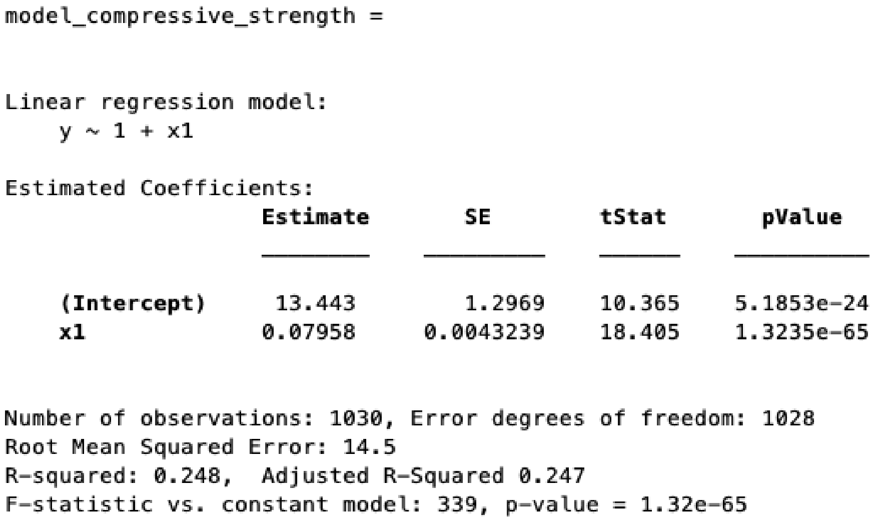


Figure 12: *Regression Model for Cement-Based Concrete*



```
model_compressive_strength = 


Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE         tStat       pValue  
                   ________    _________    ______    __________

    (Intercept)     13.443        1.2969    10.365    5.1853e-24
    x1             0.07958     0.0043239    18.405    1.3235e-65


Number of observations: 1030, Error degrees of freedom: 1028
Root Mean Squared Error: 14.5
R-squared: 0.248,  Adjusted R-Squared 0.247
F-statistic vs. constant model: 339, p-value = 1.32e-65
```

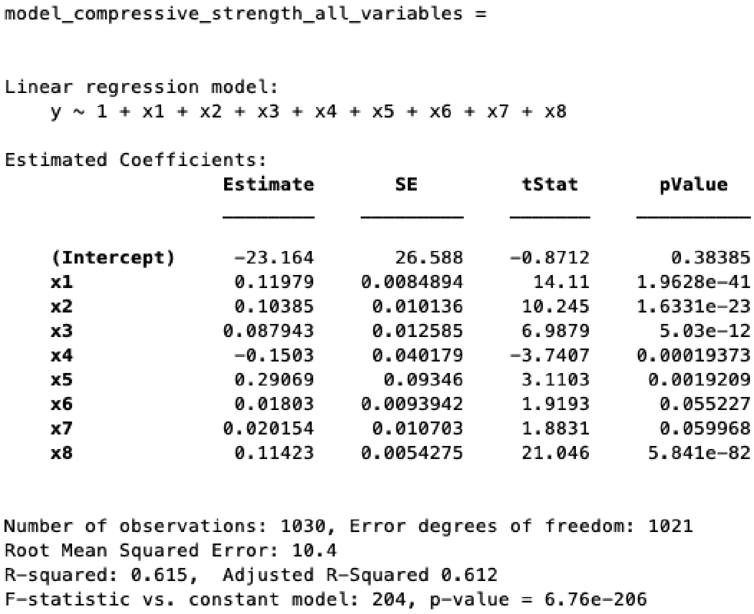
From the above snapshot it can be seen that as the coefficient of determination value is considerably low and this indicates that using a single indicator is not the best approach to predicting compressive strength. Furthermore, from Figure 12 it can be seen the predicted model significantly underestimates the actual data.

The following will expand on the above model but will now include all variables to determine the best indicates and resulting model.

1st Iteration of the Model.

Using:

- Cement-Based Component
- Blast Furnace Slag Component
- Ash Fly Component
- Water Based Component
- Super Plasticizer
- Coarse Aggregate Component
- Fine Aggregate Component
- Age Day Component



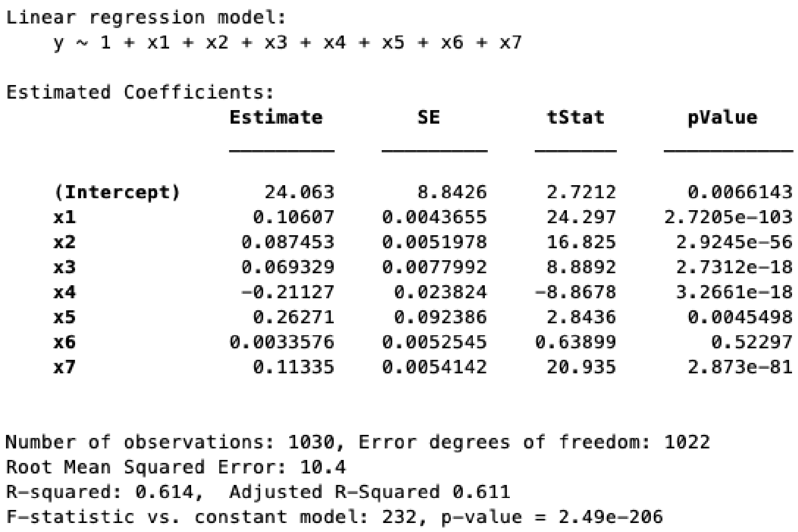
```
model_compressive_strength_all_variables =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

Estimated Coefficients:
                   Estimate       SE        tStat       pValue   
                   ________    _________    _______    __________

    (Intercept)     -23.164       26.588    -0.8712       0.38385
    x1              0.11979    0.0084894      14.11    1.9628e-41
    x2              0.10385     0.010136     10.245    1.6331e-23
    x3             0.087943     0.012585     6.9879      5.03e-12
    x4              -0.1503     0.040179    -3.7407    0.00019373
    x5              0.29069      0.09346     3.1103     0.0019209
    x6              0.01803    0.0093942     1.9193      0.055227
    x7             0.020154     0.010703     1.8831      0.059968
    x8              0.11423    0.0054275     21.046     5.841e-82


Number of observations: 1030, Error degrees of freedom: 1021
Root Mean Squared Error: 10.4
R-squared: 0.615,  Adjusted R-Squared 0.612
F-statistic vs. constant model: 204, p-value = 6.76e-206
```

From the above model it can be seen that x7 (Fine Aggregate Component) has the highest pValue and therefore needs to be removed for the second iteration of the model.

2nd Iteration of the Model.

Using:

- Cement-Based Component
- Blast Furnace Slag Component
- Ash Fly Component
- Water Based Component
- Super Plasticizer
- Coarse Aggregate Component
- Age Day Component



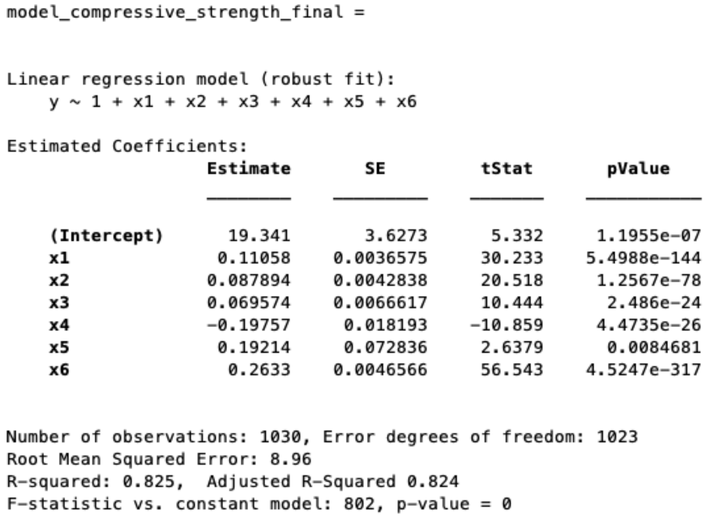
```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                   Estimate        SE         tStat       pValue   
                   _________    _________    _______    ___________

    (Intercept)       24.063       8.8426     2.7212      0.0066143
    x1               0.10607    0.0043655     24.297    2.7205e-103
    x2              0.087453    0.0051978     16.825     2.9245e-56
    x3              0.069329    0.0077992     8.8892     2.7312e-18
    x4              -0.21127     0.023824    -8.8678     3.2661e-18
    x5               0.26271     0.092386     2.8436      0.0045498
    x6             0.0033576    0.0052545    0.63899        0.52297
    x7               0.11335    0.0054142     20.935      2.873e-81


Number of observations: 1030, Error degrees of freedom: 1022
Root Mean Squared Error: 10.4
R-squared: 0.614,  Adjusted R-Squared 0.611
F-statistic vs. constant model: 232, p-value = 2.49e-206
```

The new model now shows that x6 (Coarse Aggregate Component) needs to be removed for the final model.

Final Iteration of the Model.

Using:

- Cement-Based Component
- Blast Furnace Slag Component
- Ash Fly Component
- Water Based Component
- Super Plasticizer
- Age Day Component



```
model_compressive_strength_final = 


Linear regression model (robust fit):
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6

Estimated Coefficients:
                   Estimate        SE         tStat       pValue   
                   ________    _________    _______    ___________

    (Intercept)      19.341       3.6273      5.332     1.1955e-07
    x1              0.11058    0.0036575     30.233    5.4988e-144
    x2             0.087894    0.0042838     20.518     1.2567e-78
    x3             0.069574    0.0066617     10.444      2.486e-24
    x4             -0.19757     0.018193    -10.859     4.4735e-26
    x5              0.19214     0.072836     2.6379      0.0084681
    x6               0.2633    0.0046566     56.543    4.5247e-317


Number of observations: 1030, Error degrees of freedom: 1023
Root Mean Squared Error: 8.96
R-squared: 0.825,  Adjusted R-Squared 0.824
F-statistic vs. constant model: 802, p-value = 0
```

The above snapshot shows the final model. The pValue are all sufficient and the R-squared values show a strong model. This R-squared value has increased significantly from the single-indicator-model (0.248).

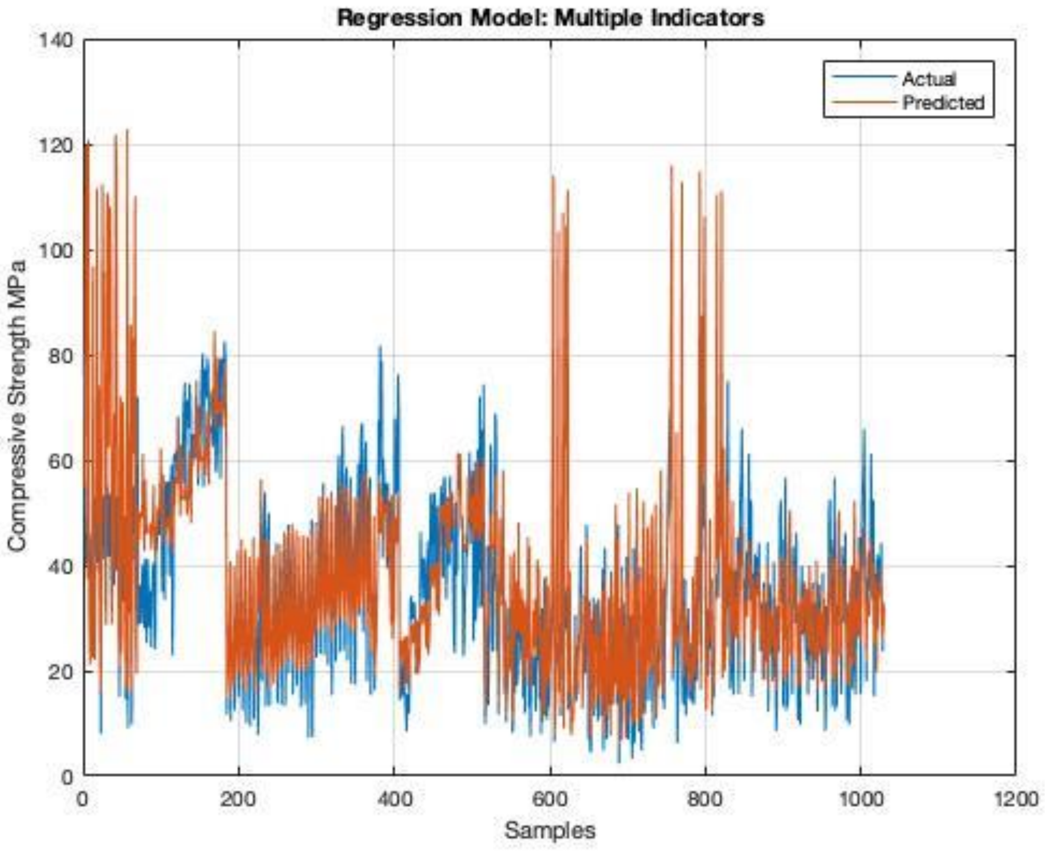


Figure 13: *Regression Model for Multiple Indicators*

Figure 13 further demonstrate the increased accuracy in predicting the compressive strength. This model has less underestimates than the previous model, however the model overestimates compressive strength readings for some of the data.

Question 5: Predicting Missing Bikeway Data

Brisbane City Council (BCC) is considering upgrades to the bikeway networks. They are using data they have gathered from sensors placed along the bike paths, which record the number of cyclists, to plan the upgrades, however there have been a number of sensor failures which have resulted in their dataset missing a number of entries. BCC have requested that you investigate if it possible to predict missing data from data gathered from other sensors on the bike path network.

You have been provided with three years' worth of data (Bike-Ped-Auto-Counts-2014.csv, Bike-Ped-Auto-Counts-2015.csv, and Bike-Ped-Auto-Counts-2016.csv), and the corresponding three years of weather data (the files named IDCJAC00XX_040913_201X_Data.csv). As an initial investigation, you have been asked to consider only these data series from the bikeway data:

- BicentennialBikewayCyclistsInbound
- GoBetweenBridgeCyclistsInbound
- KangarooPointBikewayCyclistsInbound
- NormanParkCyclistsInbound
- RiverwalkCyclistsInbound
- StoryBridgeWestCyclistsInbound

Using the three years data, you are to:

- Determine which counters are best suited to predicting the missing data in others (i.e. which, if any counters, could be used to predict BicentennialBikewayCyclistsInbound).
- Investigate if weather data can be used to help support this prediction, and if so, indicate what weather data is most helpful.
- Predict missing data where appropriate to generate a more "complete" database.
- Comment on the resulting corrected dataset. In particular:
    - What problems, if any, may arise from this approach?
    - How effective has this been in reducing missing data?
    - How trustworthy are the predicted values?

You should draw on the unit content concerning correlation and regression to answer this question. Note that you are not expected to use training/validation/testing data splits, although you are welcome to do so. No marks will be lost/gained for using/not using data splits.

(Provide any code or visualisations you use to justify your response.)

The following writing will be used to assists the Brisbane City Council in determining bikeway path upgrades. Data from the years 2014, 2015 and 2016 will be used, with specific focus on the following areas; BicentennialBikewayCyclistsInbound, GoBetweenBridgeCyclistsInbound, KangarooPointBikewayCyclistsInbound, NormanParkCyclistsInbound, RiverwalkCyclistsInbound and StoryBridgeWestCyclistsInbound. This data will be used to determine the most suitable counters to predict the missing data, investigate if weather can provide more accuracy to the model and consequently, predict missing data to generate a "complete" database.

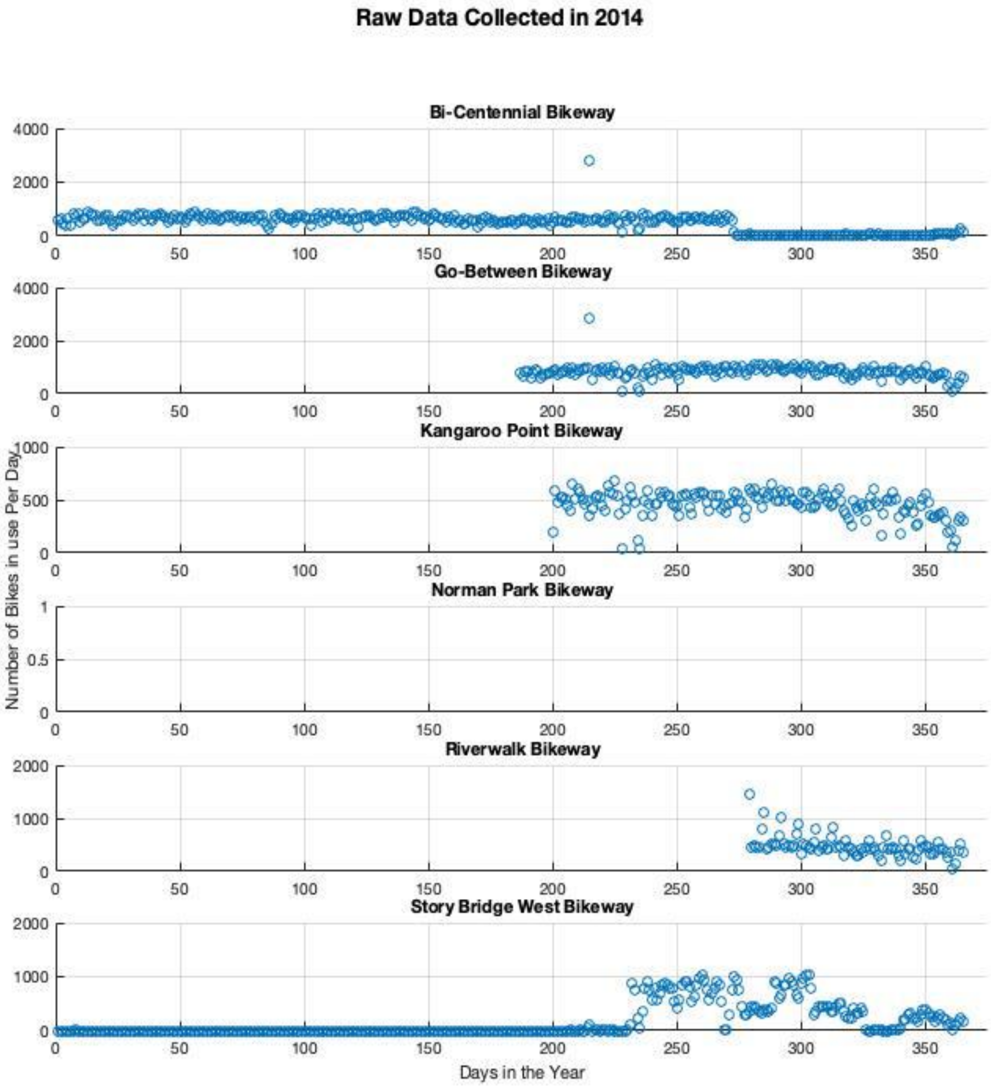


Figure 14: *Raw Data Collected for the Bike paths in 2014*

- **Data Observations**:
  - Bi-Centennial Bikeway
    - From the graph it appears all data is present
  - Go-Between Bikeway
    - Data is missing for approximately the first half of the year (day 1 - ~180)
    - A significant outlier is also seen at ~day 220
  - Kangaroo Point Bikeway
    - Data shows a significant similarity to the Go-Between data
  - Norman Park Bikeway
    - All readings are missing
  - Riverwalk Bikeway
    - A significant amount of data is missing (~275 days missing)
  - Story Bridge West Bikeway
    - Data for the first ~225 days boarders or is 0, this will need to be investigated as it may represent missing data

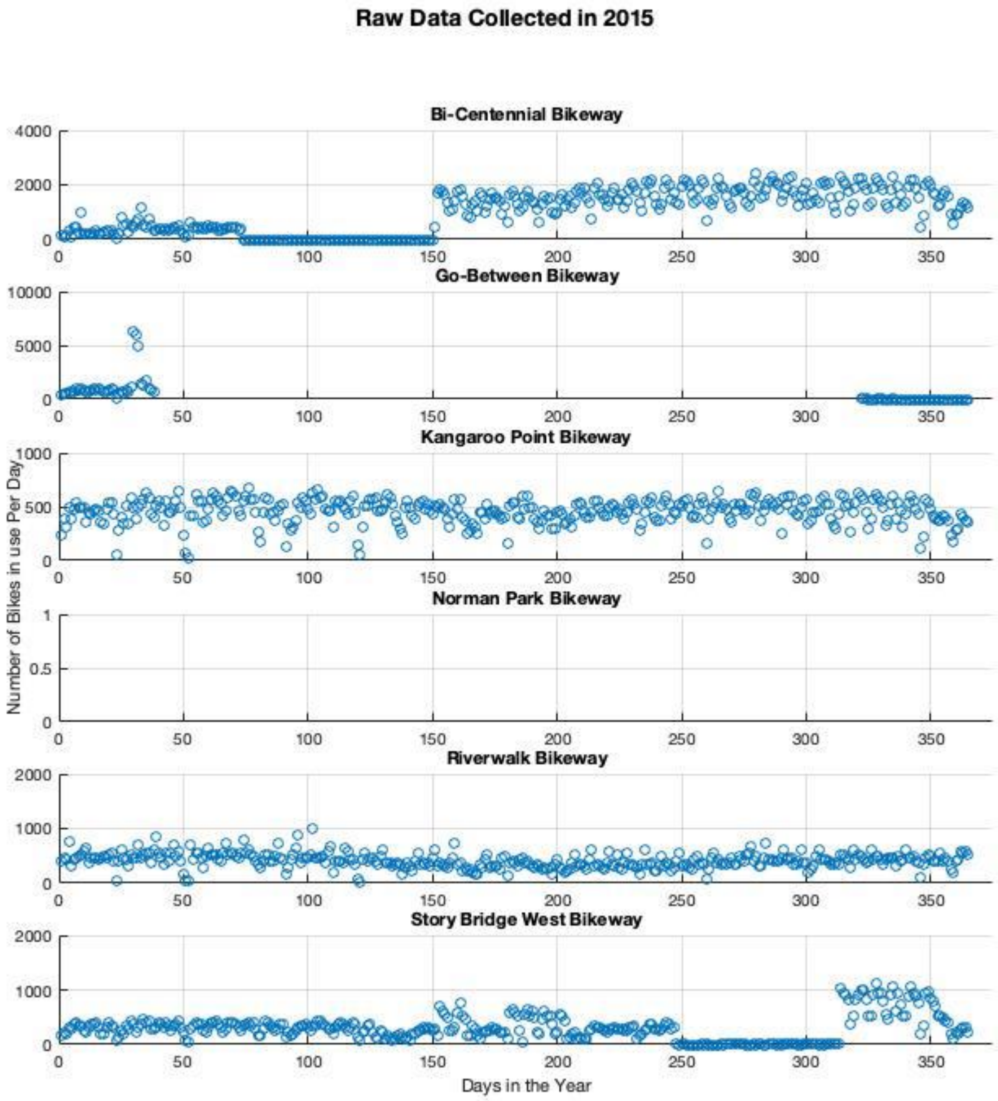


Figure 15: *Raw Data Collected for the Bike paths in 2015*

- **Data Observations**:
    - Bi-Centennial Bikeway
        - Erratic readings until ~day 150, from here it stabilises
    - Go-Between Bikeway
        - Significant missing data, from days ~40 - ~315
    - Kangaroo Point Bikeway
        - Appears to be no missing data and a well collected data set
    - Norman Park Bikeway
        - All readings are missing
    - Riverwalk Bikeway
        - Appears to be no missing data and a well collected data set
    - Story Bridge West Bikeway
        - Readings for day ~245 - ~315 appear to boarder or are 0's, this will need to be investigated as it may represent missing data

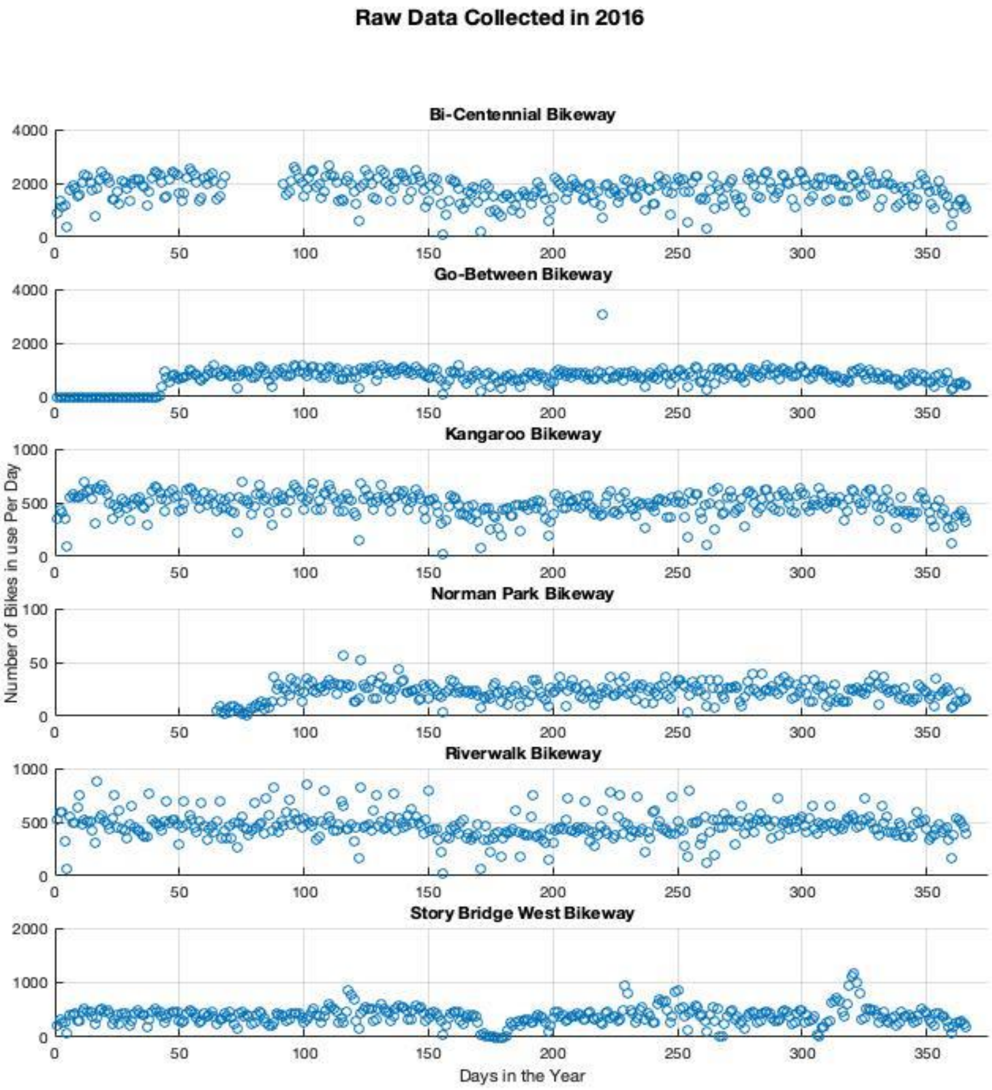


Figure 16: *Raw Data Collected for the Bike paths in 2016*

- **Data Observations**:
  - Bi-Centennial Bikeway
    - Missing data from ~day 65 - 90
  - Go-Between Bikeway
    - Data for the first ~40 days appears to be 0’s, this may represent missing data, all other readings appear to be acceptable
  - Kangaroo Point Bikeway
    - Appears to be no missing data and a well collected data set
  - Norman Park Bikeway
    - Missing data for the first ~60 days
  - Riverwalk Bikeway
    - Appears to be no missing data and a well collected data set
  - Story Bridge West Bikeway
    - Appears to be no missing data and a well collected data set

Part A: Determining the most suitable counters to predicting missing data from Figures 14, 15 and 16 and correlation.

2014:

From Figure 14 and Table 6 it can be seen that the most suitable counter is *Bi-Centennial.* The Go-Between, Kangaroo Point and Riverwalk Bikeways had no data for the ~first half – ¾ of the year and therefore are unable to provide any accuracy to the model. The Story Bridge West Bikeway had 0's for the ~first half of the year and therefore this data would be unable to model any other bikeways. Norman Park was not selected as there is no data, therefore there is no opportunity for missing data to be modelled after collected data. Riverwalk was identified as an unsuitable counter as the missing data was considerable.

Table 6: *Correlation Data for the Sample Bikeways: 2014*

	Bi-Centennial	**Go-Between**	**Kangaroo Point**	**Norman Park**	**Riverwalk**	**Story Bridge West**
Bi-Centennial	1.0000	0.3899	0.1668	NaN	-0.0488	-0.3571
Go-Between	0.3899	1.0000	0.5599	NaN	0.4515	0.2473
Kangaroo Point	0.1668	0.5599	1.0000	NaN	0.4902	0.3455
Norman Park	NaN	NaN	NaN	NaN	NaN	NaN
Riverwalk	-0.0488	0.4515	0.4902	NaN	1.0000	0.2419
Story Bridge West	-0.3571	0.2473	0.3455	NaN	0.2419	1.0000

2015:

From Figure 15 and Table 7 it can be seen that the most suitable counters are *Kangaroo Point* and *Riverwalk*. The Bi-Centennial and Story Bridge West Bikeways both had data that would not allow a model to be created accurately. Norman Park was not selected as there is no data, therefore there is no opportunity for missing data to be modelled after collected data. The sample size of Riverwalk is too small to provide confidently predicted data.

Table 7: *Correlation Data for the Sample Bikeways: 2015*

	Bi-Centennial	**Go-Between**	**Kangaroo Point**	**Norman Park**	**Riverwalk**	**Story Bridge West**
Bi-Centennial	1.0000	-0.3831	0.2110	NaN	-0.1544	0.1297
Go-Between	-0.3831	1.0000	0.1417	NaN	0.2304	-0.2941
Kangaroo Point	0.2110	0.1417	1.0000	NaN	0.4993	0.2566
Norman Park	NaN	NaN	NaN	NaN	NaN	NaN

Riverwalk	-0.1544	0.2304	0.4993	NaN	1.0000	0.1575
Story Bridge West	0.1297	-0.2941	0.2566	NaN	0.1575	1.0000

2016:

From Figure 16 and Table 8 it can be seen that the readings collected in 2016 provide the most consistent data and therefore will be able to produce the most accurate predicted data. All bikeways were identified are suitable predicters. Both sample sizes and consistency were acceptable for all sites.

Table 8: *Correlation Data for the Sample Bikeways: 2016*

	Bi-Centennial	**Go-Between**	**Kangaroo Point**	**Norman Park**	**Riverwalk**	**Story Bridge West**
Bi-Centennial	1.0000	0.2938	0.8741	0.5232	0.2543	0.6627
Go-Between	0.2938	1.0000	0.2935	0.3165	0.1446	0.2174
Kangaroo Point	0.8741	0.2935	1.0000	0.4383	0.5651	0.5992
Norman Park	0.5232	0.3165	0.4383	1.0000	0.2716	0.3199
Riverwalk	0.2543	0.1446	0.5651	0.2716	1.0000	0.1701
Story Bridge West	0.6627	0.2174	0.5922	0.3199	0.1701	1.0000

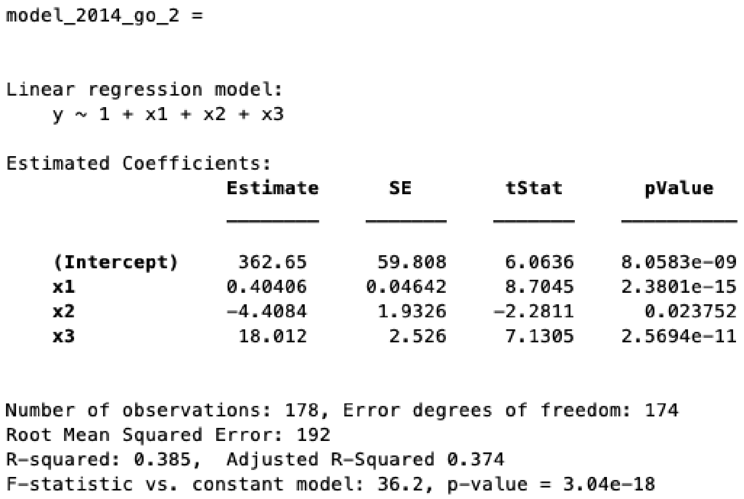
Part B & C: Investigating if weather data can be used to support the above and predicting a more “complete” database.

2014:

The following will show the predicted data for Bi-Centennial, Go-Between and Kangaroo Point Bikeways. No predicted data was able to be generated for:

- Bi-Centennial – 0’s assumed to be representing missing data, however cannot be removed
- Norman Park – no data was given therefore none can be predicted
- Riverwalk – data set was too small
- Story Bridge West Bikeways – missing data was presented as 0’s, making it insufficient

Go-Between Bikeway Predictive Model:



```
model_2014_go_2 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue  
                   ________    _______    _______    __________

    (Intercept)     362.65      59.808     6.0636    8.0583e-09
    x1             0.40406     0.04642     8.7045    2.3801e-15
    x2             -4.4084      1.9326    -2.2811      0.023752
    x3              18.012       2.526     7.1305    2.5694e-11


Number of observations: 178, Error degrees of freedom: 174
Root Mean Squared Error: 192
R-squared: 0.385,  Adjusted R-Squared 0.374
F-statistic vs. constant model: 36.2, p-value = 3.04e-18
```

The snapshot above shows the results of the model used to predict the missing data in 2014 on the Go-Between data set. From such it can be seen that the pValues are all acceptable, however, the R-squared value is considerably lower than that in the above model. The model utilised the following bikeway and weather counters: Bi-Centennial, rainfall and solar exposure (x1, x2 & x3, respective).

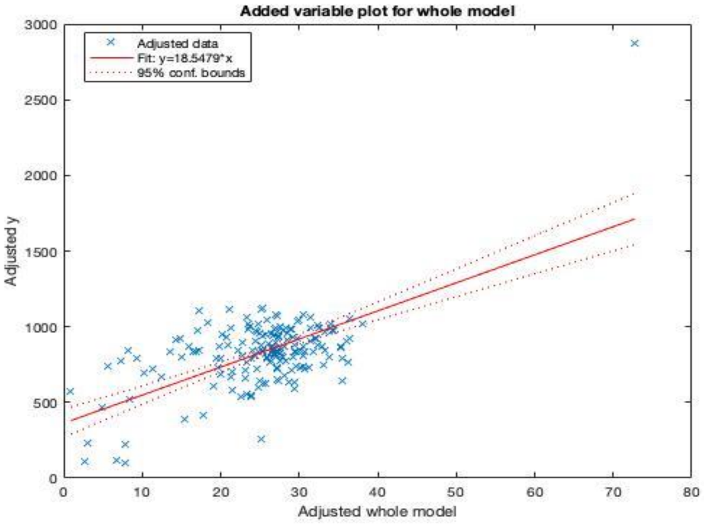


Figure 17: *Final Model Predictive Plot (model_2014_go_2)*

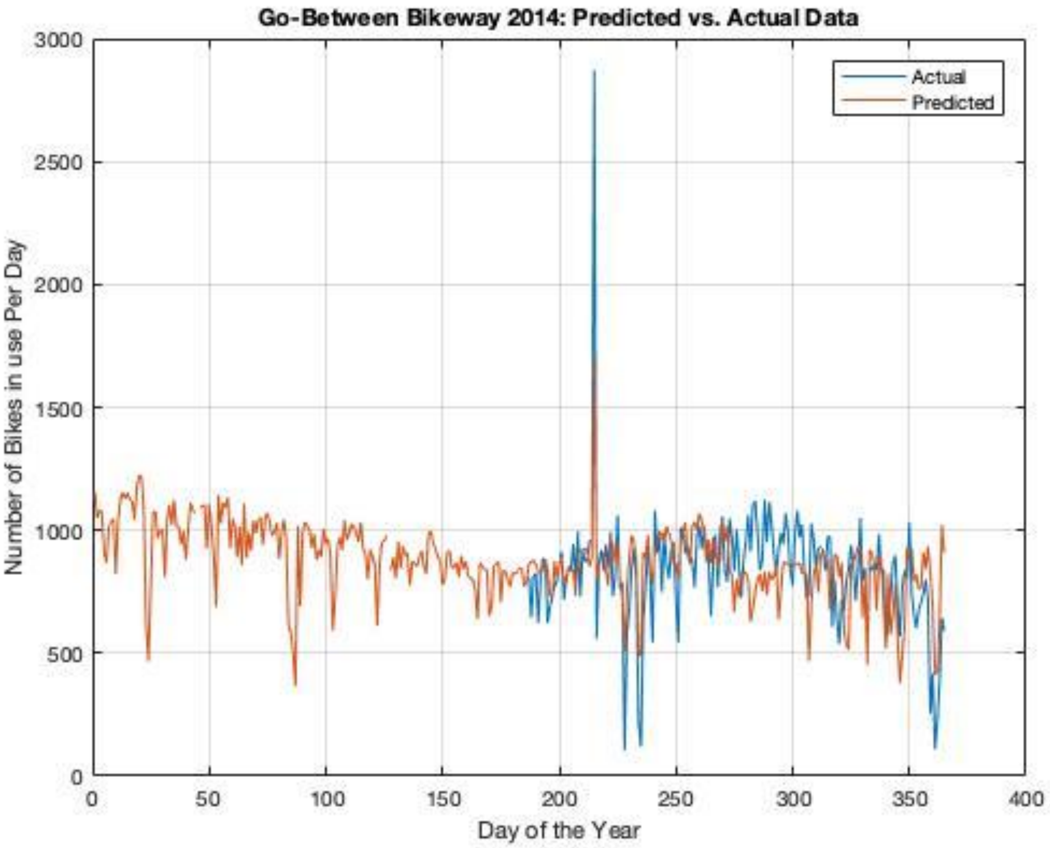
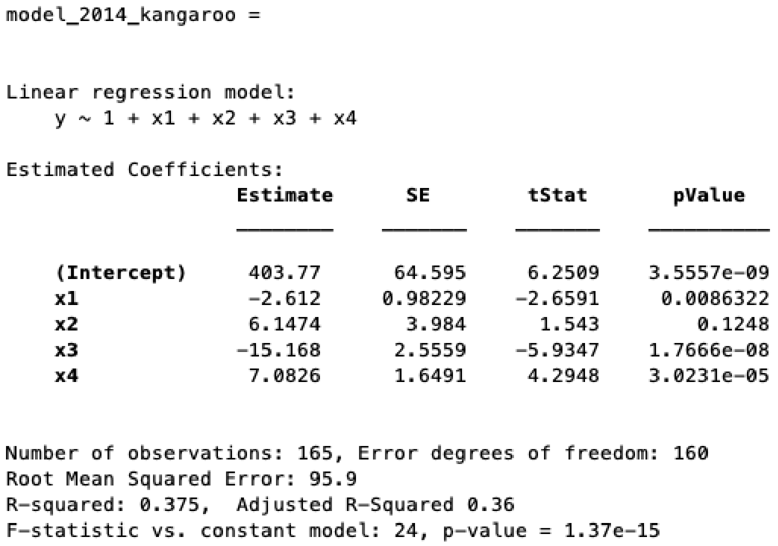


Figure 18: *Predicted vs. Actual data for the Go-Between Bikeway in 2014*

Kangaroo Point Bikeway Predictive Model:



```
model_2014_kangaroo =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)     403.77      64.595     6.2509    3.5557e-09
    x1              -2.612     0.98229    -2.6591     0.0086322
    x2              6.1474       3.984      1.543        0.1248
    x3             -15.168      2.5559    -5.9347    1.7666e-08
    x4              7.0826      1.6491     4.2948    3.0231e-05


Number of observations: 165, Error degrees of freedom: 160
Root Mean Squared Error: 95.9
R-squared: 0.375,  Adjusted R-Squared 0.36
F-statistic vs. constant model: 24, p-value = 1.37e-15
```

The snapshot above shows the results of the model used to predict the missing data in 2014 on the Kangaroo Point data set. From such it can be seen that the pValues are all acceptable, however the R-squared value is low. The model utilised the following weather counters: rainfall, maximum temperature, minimum temperature and solar exposure (x1, x2, x3 & x4, respective).

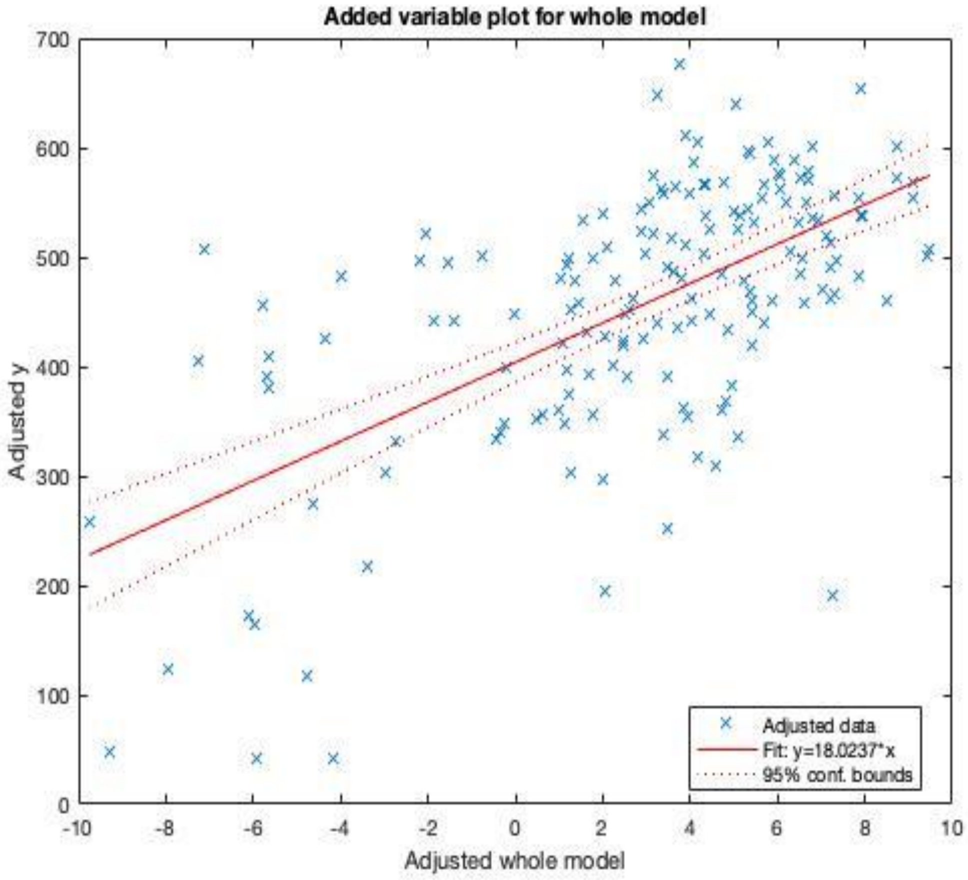


Figure 19: *Final Model Predictive Plot (model_2014_kangaroo)*

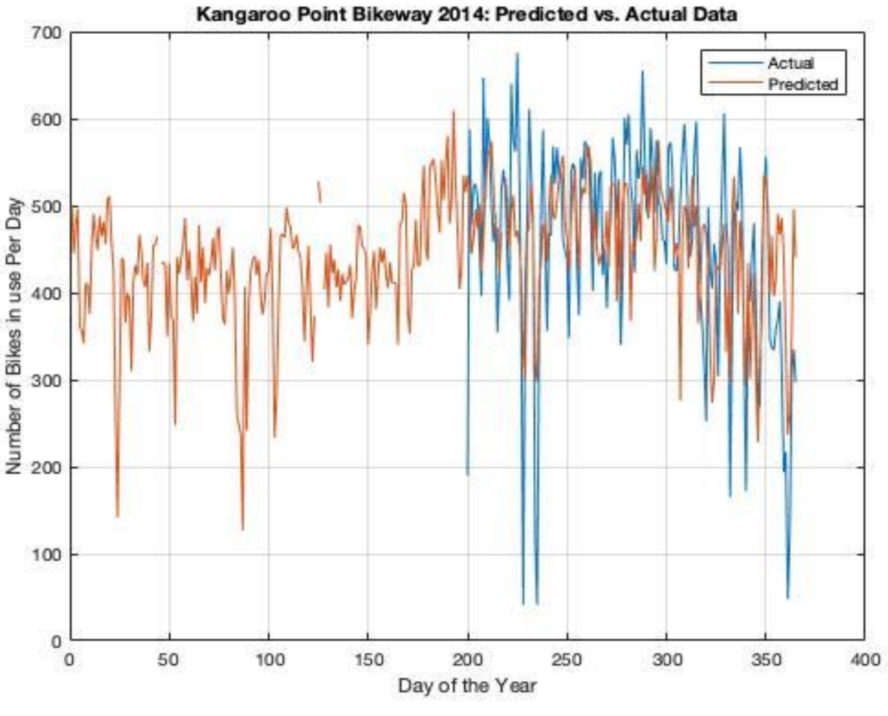


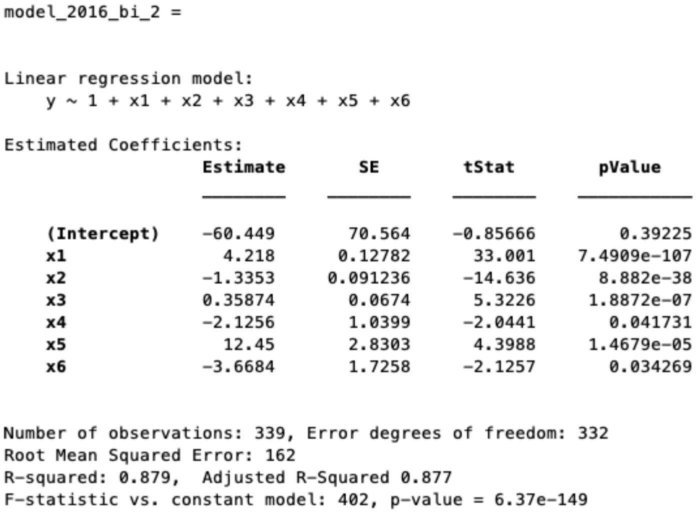
Figure 20: *Predicted vs. Actual data for the Bi-Centennial Bikeway in 2014*

2015:

See **Part D** for the explanation as to why there is no predictive models for 2015 data.

2016:

Bi-Centennial Bikeway Predictive Model:



```
model_2016_bi_2 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    ________    __________

    (Intercept)     -60.449      70.564    -0.85666       0.39225
    x1                4.218     0.12782      33.001    7.4909e-107
    x2              -1.3353    0.091236     -14.636      8.882e-38
    x3              0.35874      0.0674      5.3226     1.8872e-07
    x4              -2.1256      1.0399     -2.0441       0.041731
    x5                12.45      2.8303      4.3988     1.4679e-05
    x6              -3.6684      1.7258     -2.1257       0.034269


Number of observations: 339, Error degrees of freedom: 332
Root Mean Squared Error: 162
R-squared: 0.879,  Adjusted R-Squared 0.877
F-statistic vs. constant model: 402, p-value = 6.37e-149
```

The snapshot above shows the results of the model used to predict the missing data in 2016 on the Bi-centennial data set. From such it can be seen that the pValues are all acceptable and the R-squared value shows high accuracy in the model. The model utilised the following bikeway and weather counters: Kangaroo Point, Riverwalk, Story Bridge West, rainfall, maximum temperature and solar exposure (x1, x2, x3, x4, x5 & x6, respective).

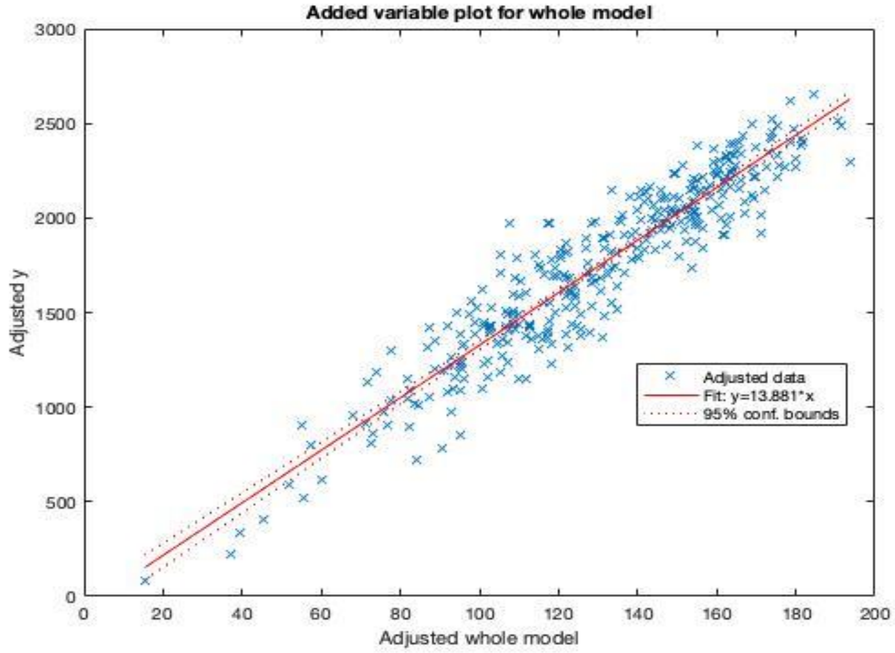


Figure 21: *Final Model Predictive Plot (model_2016_bi_2)*

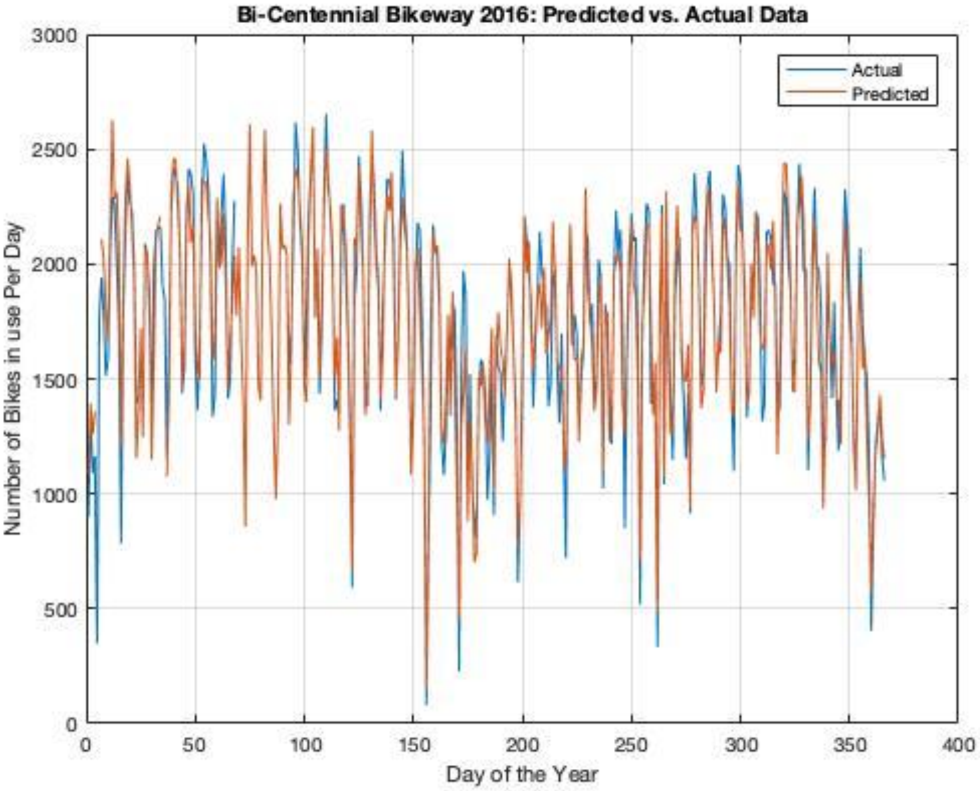
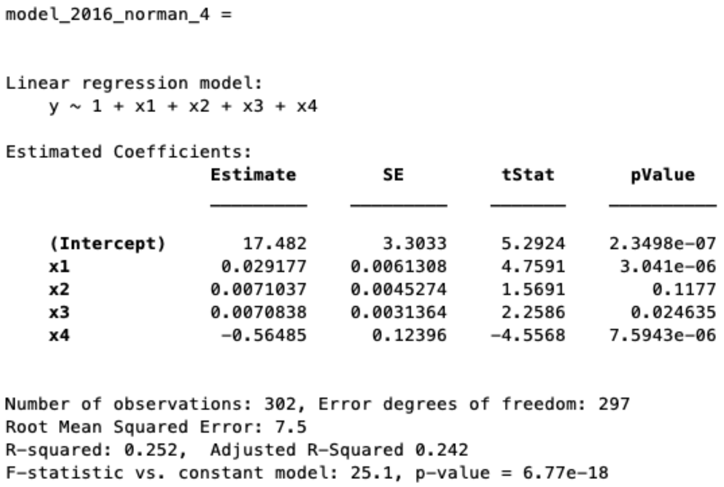


Figure 22: *Predicted vs. Actual data for the Bi-Centennial Bikeway in 2016*

Norman Park Bikeway Predictive Model:



```
model_2016_norman_4 =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate        SE         tStat       pValue  
                   _________    _________    _______    __________

    (Intercept)       17.482       3.3033     5.2924    2.3498e-07
    x1              0.029177    0.0061308     4.7591     3.041e-06
    x2             0.0071037    0.0045274     1.5691        0.1177
    x3             0.0070838    0.0031364     2.2586      0.024635
    x4              -0.56485      0.12396    -4.5568    7.5943e-06


Number of observations: 302, Error degrees of freedom: 297
Root Mean Squared Error: 7.5
R-squared: 0.252,  Adjusted R-Squared 0.242
F-statistic vs. constant model: 25.1, p-value = 6.77e-18
```

The snapshot above shows the results of the model used to predict the missing data in 2014 on the Bi-centennial data set. From such it can be seen that the pValues are all acceptable, however, the R-squared value is low. The model utilised the following bikeway and weather counters: Kangaroo Point, Riverwalk and Story Bridge West and maximum temperature (x3 & x4, respective).

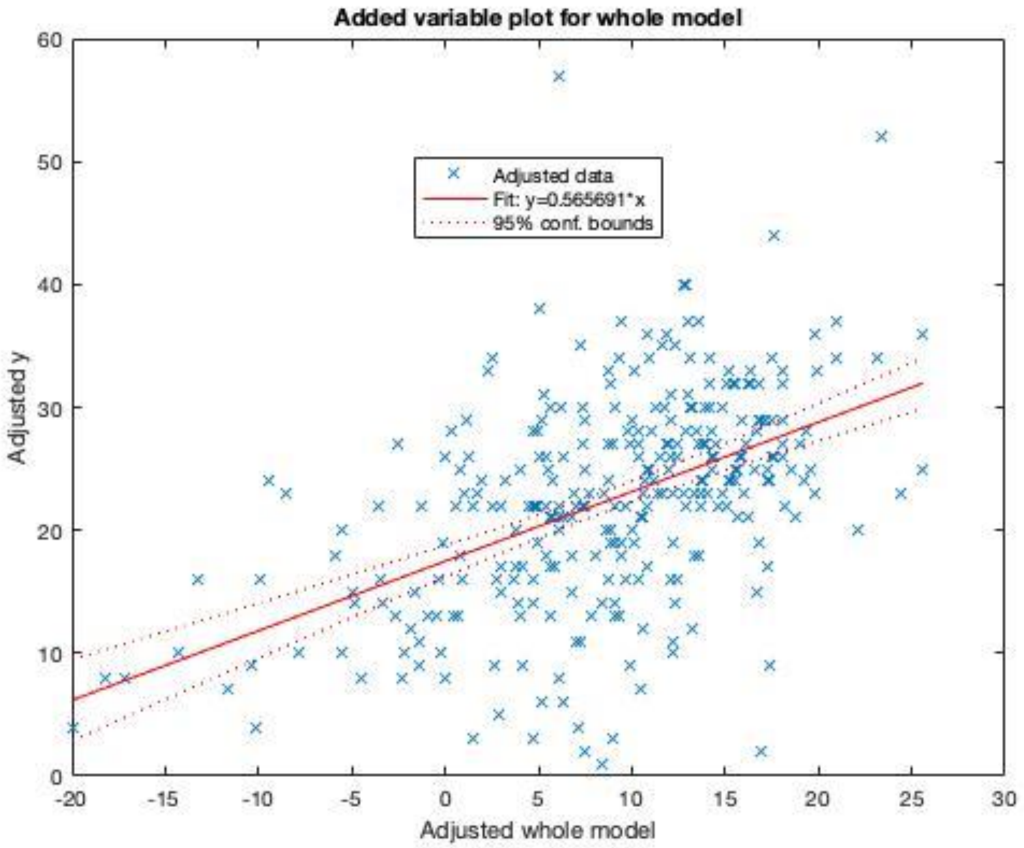


Figure 23: *Final Model Predictive Plot (model_2016_norman_4)*

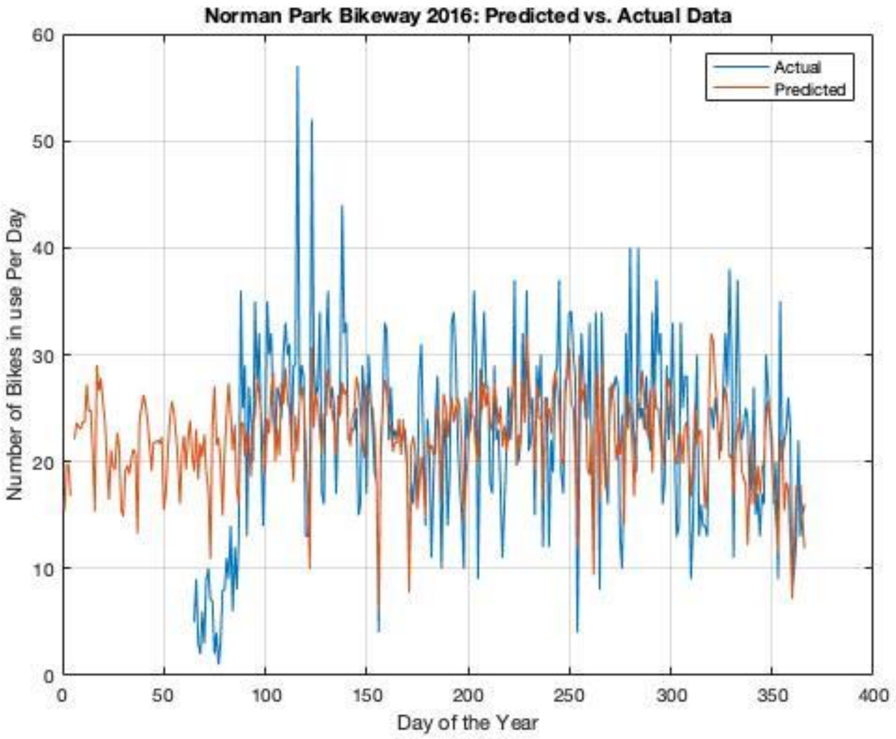


Figure 24: *Predicted vs. Actual data for the Norman Park Bikeway in 2016*

Part D: Comment on the result corrected datasheet. In particular:

- What problems, if any, may arise from this approach?

0's Representing Missing Data: No predictive models were able to be made for the 2015 database due as the two subject counters to be modelled (Bi-Centennial and Story Bridge West) had insufficient missing data. The 0's were assumed to represent missing data, however as this is an assumption, they cannot be removed in order to model predictive behaviour. Therefore, if a data set is to be modelled and it contains 0's (missing data) then these should be removed.

Small Sample Sizes: The size of the data also contributed to reducing accuracy, that is, the databases with small sample sizes did not allow for an accurate model to be produced. This was seen in 2014 – Riverwalk Bikeway and 2015 – Go-Between Bikeway.

- How effective has this been in reducing missing data?

Figure 24 and the prediction of missing data for the Bi-Centennial Bikeway 2016 can be seen to considerably effective in reducing missing data. The results presented demonstrated that the model was able to accurately cover the missing data and fit to the actual data. Furthermore, the model produced an R-squared value of 0.877.

- How trustworthy are the predicted values?

The trustworthiness of the predicted models is predicated on both the sample size of the original data and size of the counters data. The R-squared value should also be as close to one as possible. In the models presented above it can be seen that the modelling of the Bi-Centennial Bikeway 2016 is the most trustworthy, then

Bibliography

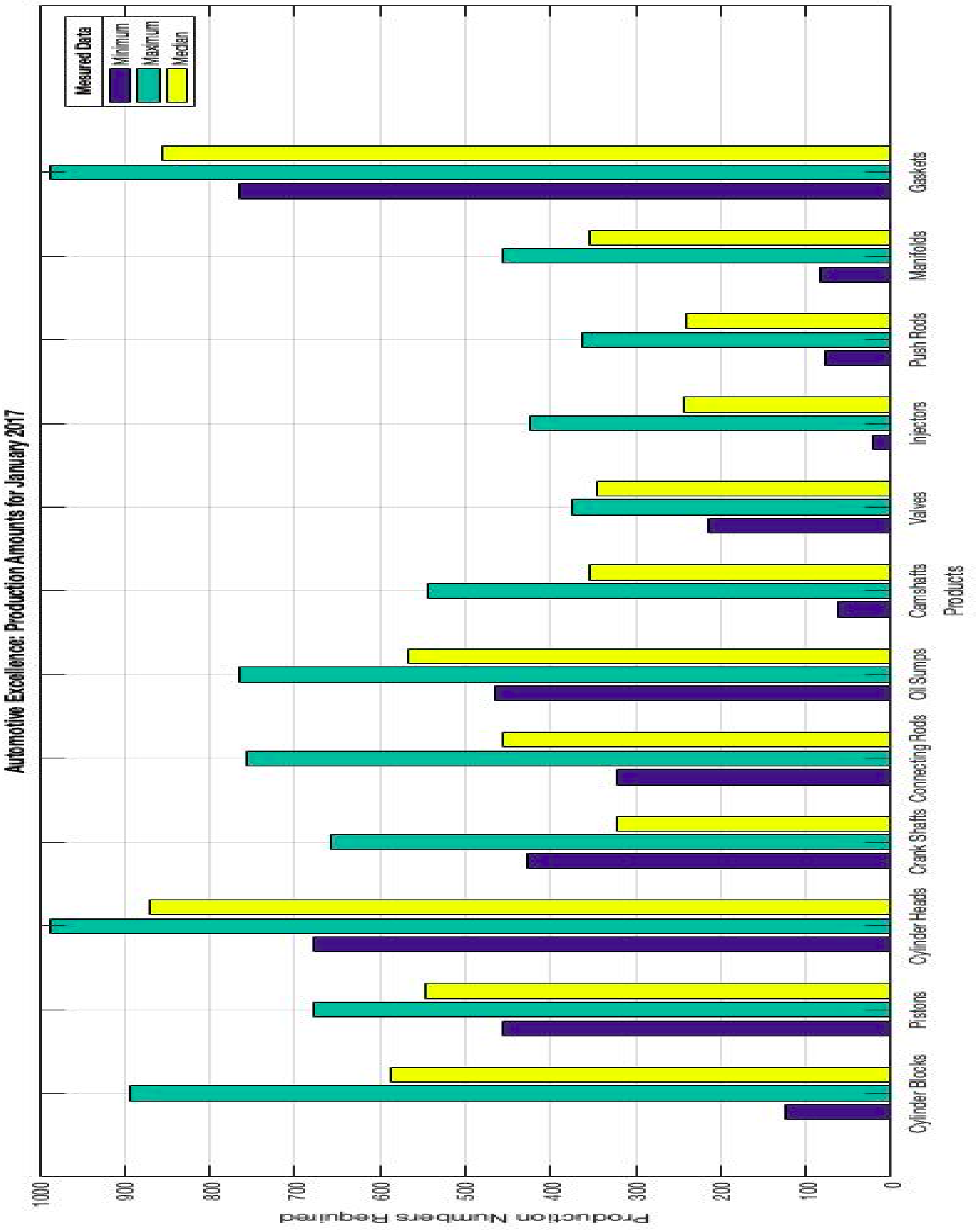
Spelke, E., 2004. *Principles of Object Perception.* [Online]

Available at: https://doi.org/10.1016/0364-0213(90)90025-R

[Accessed 15 8 2019].

Appendices

Appendix A:



How will my work be graded?

Your responses to questions will be marked using the rubric set out below.

All questions are worth equal marks.

Passing grades

Grade 7 work…

- Clear and concise analysis and visualisation (where required) of given data.
- This level of work implies excellence in thinking.
- It is on the whole clear, precise and well-reasoned, and all aspects of the problems are discussed.
- Terms and distinctions are used effectively.
- Work at this level demonstrates a mind beginning to take charge of its own ideas, assumptions, inferences and intellectual processes.

Grade 6 work…

…demonstrates more strength than weaknesses and is more consistent in high-level performance than grade 5 work. It has some distinctive weaknesses, though no major ones

- Analysis is mostly clear, precise and well-reasoned/ visualisation of given data can be improved
- This level of work implies excellence in thinking.
- Terms and distinctions are used effectively.
- Work at this level demonstrates a mind beginning to take charge of its own ideas, assumptions, inferences and intellectual processes.

Grade 5 work…

… demonstrates more than a minimal level of skill, but it is also highly inconsistent, with as many weaknesses as strengths.

- Clear analysis with some flaws in methodology/some questions are unexplored/weak visualisation of data
- This level of work shows some signs of critical thinking.
- Clear analysis but occasionally lacks reasoning.
- Terms and distinctions are often used effectively.
- There is some evidence that the student is genuinely engaged in the task of taking charge of his or her thinking.

Grade 4 work…

… demonstrates only a minimal level of understanding and skill in the subject

- Minimal analysis/weak representation of given data.
- This level of work shows minimal signs of critical thinking.
- Analysis is not clear and often lacks reasoning.
- Some terms and distinctions are used effectively.

Failing grades

Grade 3 work…

…demonstrates a pattern of unskilled thinking

- Incorrect analysis/flawed methodology to analyse data/poor representation of data.
- The work suggests that the student is trying to get through the assessment by means of wrote recall, attempting to acquire knowledge by memorisation rather than through comprehension and understanding.
- The work suggests the student is not developing the skills of thought and knowledge requisite to understanding how to read and make sense of data.
- Work at this level, on close examination, typically reveals characteristics including that the student does not understand the basic nature of what it means to think within the context of the assessment, and in any case does not display the thinking skills and abilities which is at the heart of the assessment.
- The work is vague, imprecise, and unreasoned.
- There is little evidence that the student is genuinely engaged in the task of taking charge of his or her thinking.
- The work suggests the student is simply going through the motions without really putting any significant effort into thinking through the questions.
- The work suggests the student is
  - not analysing issues clearly,
  - not formulating information clearly,
  - not accurately distinguishing the relevant from the irrelevant,
  - not identifying key questionable assumptions,
  - not clarifying key concepts,
  - not identifying relevant competing points of view,
  - not reasoning carefully from clearly stated premises,
  - not tracing implications and consequences.
- The work does not display discernible reasoning and problem-solving skills

Grade 2/1 work…

…demonstrates a pattern of unskilled thinking and fails to do the required work/minimum work done.

Grade 0 work…

… No relevant response to the questions