THE UNIVERSITY OF SYDNEY

# BUSS 6002 Data Science in Business

## Group Assignment

## Group 256

480052425 Jingxuan Liu (Daisy)

480148315 Chen Jing (Steve)

450530575 Ankur Bedi

## Part 1: Cohesive EDA and Feature Selection

*Motivation Behind Choosing Linear Regression Model*

The reason behind selecting the Linear regression Model for predicting the log RV of daily stock returns of the Commonwealth Bank of Australia is the availability of lag features in the dataset. While linear regression may not be the best modelling technique for forecasting time series data, it is a highly desirable model for lagged value predictors (Hyndman & Athanasopoulos, 2013). In general, lag values better represent the effects of policy change or for instance. It is most likely that any shift in CBA strategy will have a lagging impact on stock price volatility (log RV) (Hyndman & Athanasopoulos, 2013).

Secondly, the initial plotting of the data shows a stationary trend around the mean. This would satisfy the constant variance assumption of the linear regression model.

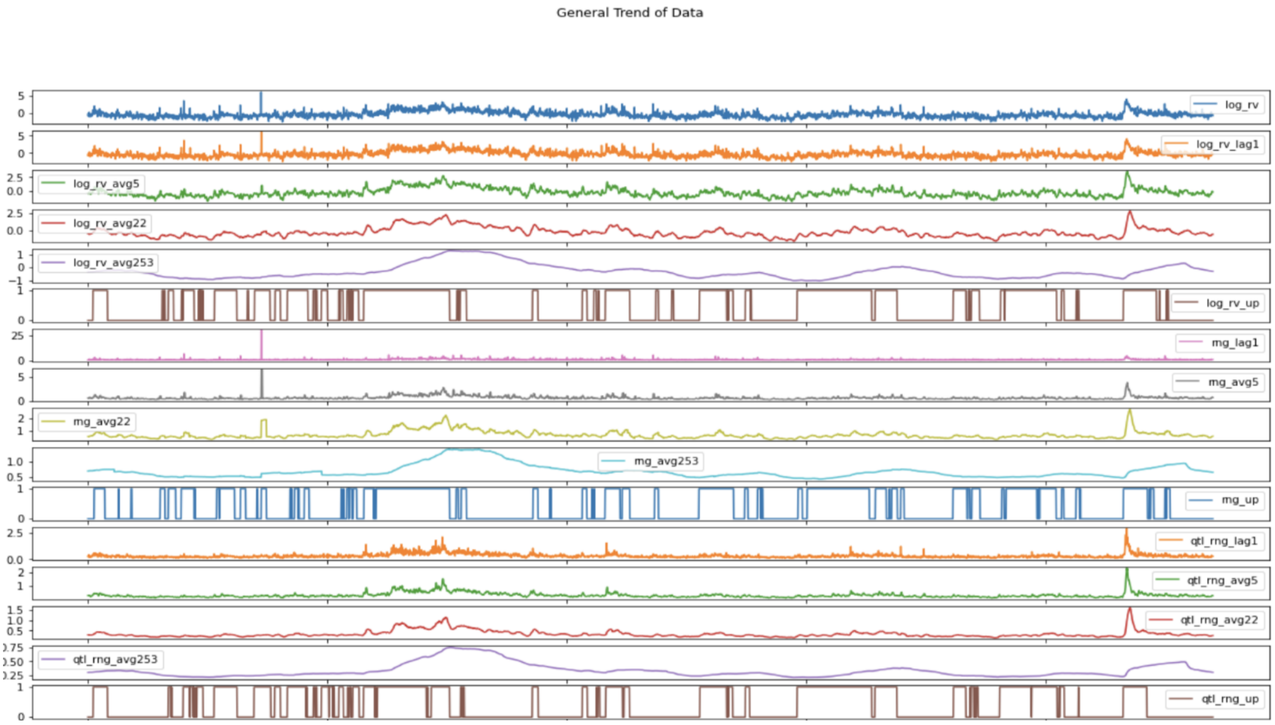


*figure 1 Time Series Plot of Entire Dataset*

The above two points are the major motivation behind using the Linear regression model, but overall having some inclination of modelling method helps in using targeted Exploratory Data Analysis (EDA) and Feature Selection Methods.

*Continuous Variable Analysis*

While selecting features that are predictive of future log RV value of CBA at least one-day-ahead. We are basing this analysis on satisfying the following three goals:

- Overfitting  - This will be achieved by selecting the correct number of features
- Underfitting - This will be achieved by choosing the right number of features
- Linear regression Assumptions - This will be the focus of EDA

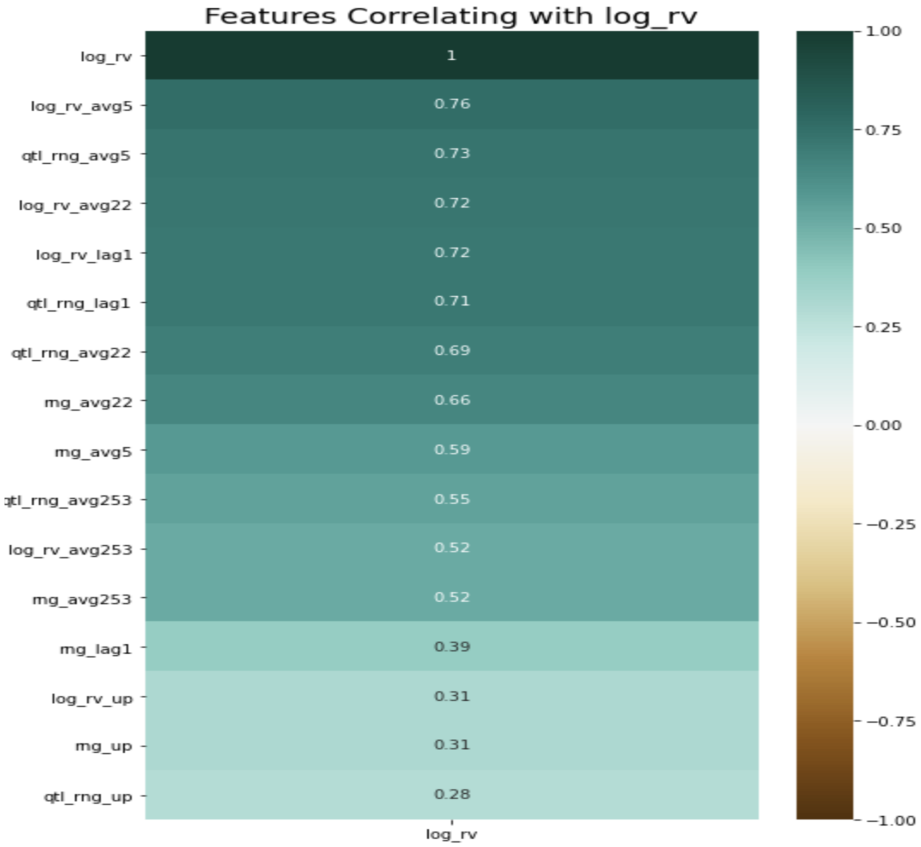


*figure 2 Features Correlation Heatmap*

As shown above, Pearson's correlation coefficient shows log_rv_avg5 being the most linearly related predictor variable with the response, followed by the qtl_rng_avg5.

While the Pearson's correlation coefficient index is crucial in selecting linear features, it is equally important to remember the other assumptions of the linear regression model, in this case multicollinearity. Hence, choosing one variable from each feature set could be an efficient way to reduce the effects of multicollinearity.

Further analysing the correlation - *features (log_rv_avg5, qtl_rng_avg5,rng_avg22)*

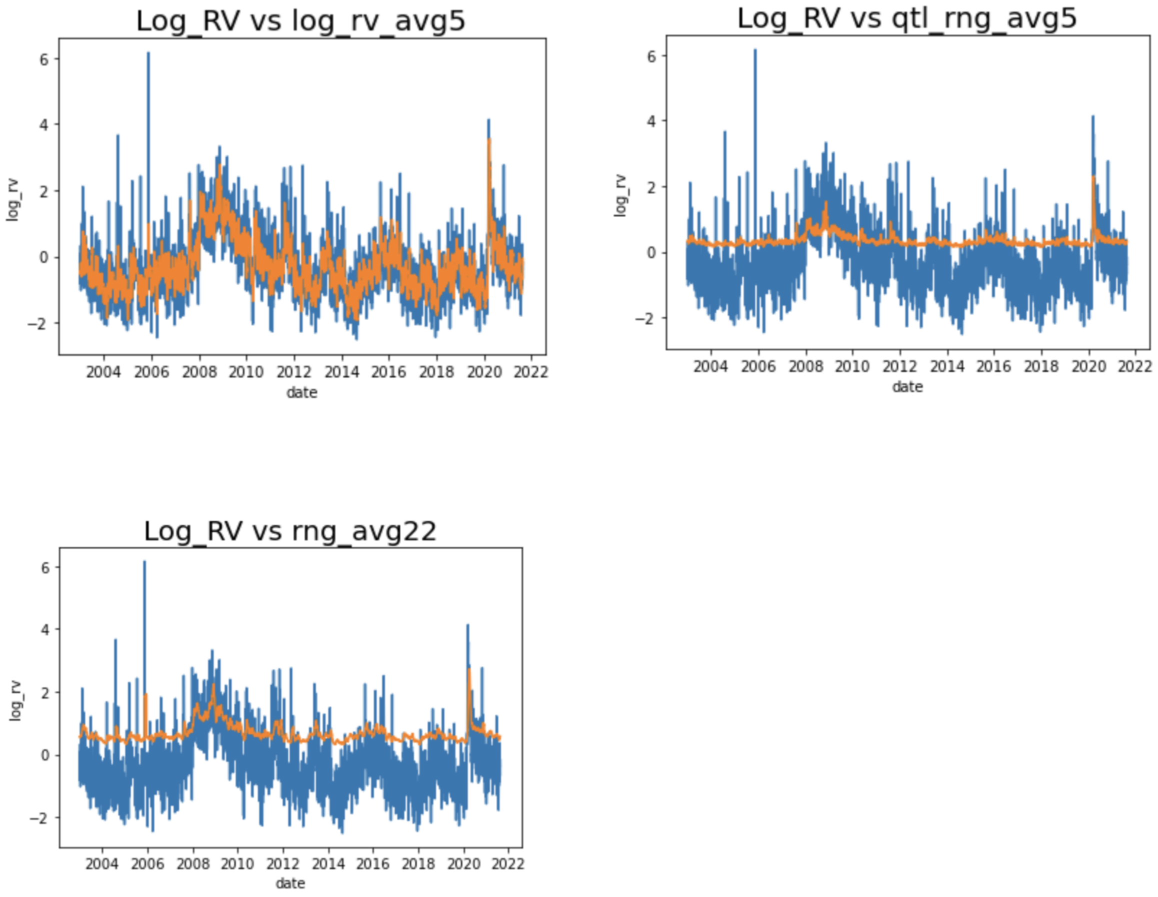


*figure 3 Further analysing the correlation - features (log_rv_avg5, qtl_rng_avg5,rng_avg22)*

*Multicollinearity*

One of the main concerns in selecting the features was the unsatisfactory multicollinearity assumption, which is a feature of lag predictors, and to what extent this will affect the accuracy of our predictions. As shown below features *(log_rv_avg5, qtl_rng_avg5,rng_avg22)* are highly correlated

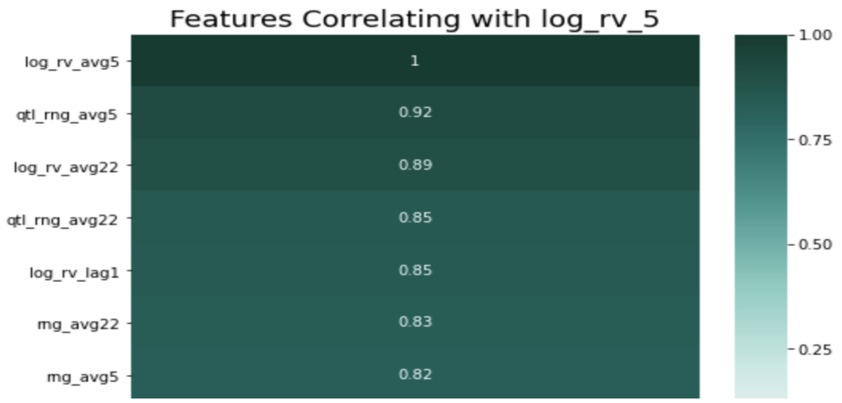


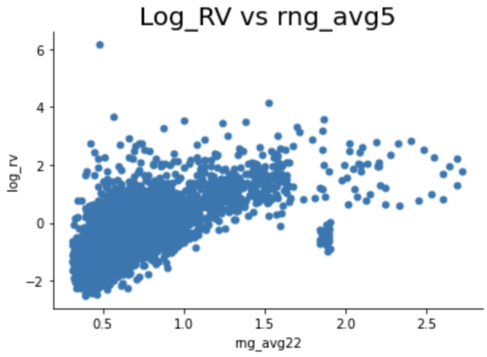
*figure 4 Features Correlation with Log_rv_5*

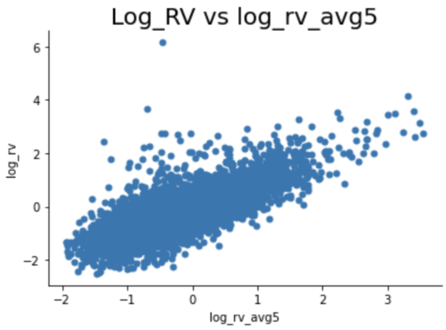
While there is multicollinearity between the variables, its effect on our prediction should be minimal as we are just interested in the final prediction of our log_rv value. Quoting an excerpt by Kurtner in Applied linear statistical Models - " *The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations"* (Kutner, 2005)

Multicollinearity is an issue in linear regression when we are interested in interpreting parameter estimates, such as predicting house prices by analysing the number of bedrooms. Here, multicollinearity will lead to inaccurate predictions. But in our case, we are only interested in the final prediction, and we can check their accuracy through a given metric.

*Linearity*

The effectiveness of the chosen features in predicting the future log RV value by applying a linear regression model is further validated by the scatter plots below, which shows a strong linear relationship between predictors and the response variable.





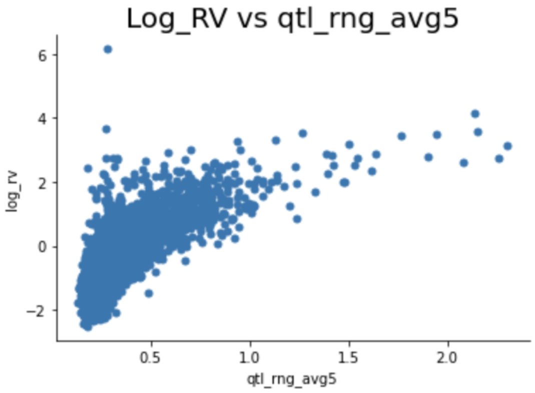


*figure 5 Scatter Plots*

### *Categorical Variable Analysis*

In order to interpret the trends of three binary variables (log_rv_up, rng_up, qtl_rng_up), three bar charts are generated. The indicator 0 represents a downward trend with avg22< avg253, while 1 indicates an upward trend with avg22 > avg253.

As illustrated in the following charts, The data in all features located in category 1 exceeds the data in category 0, which indicates that the log_rv is on a recent downward trend.

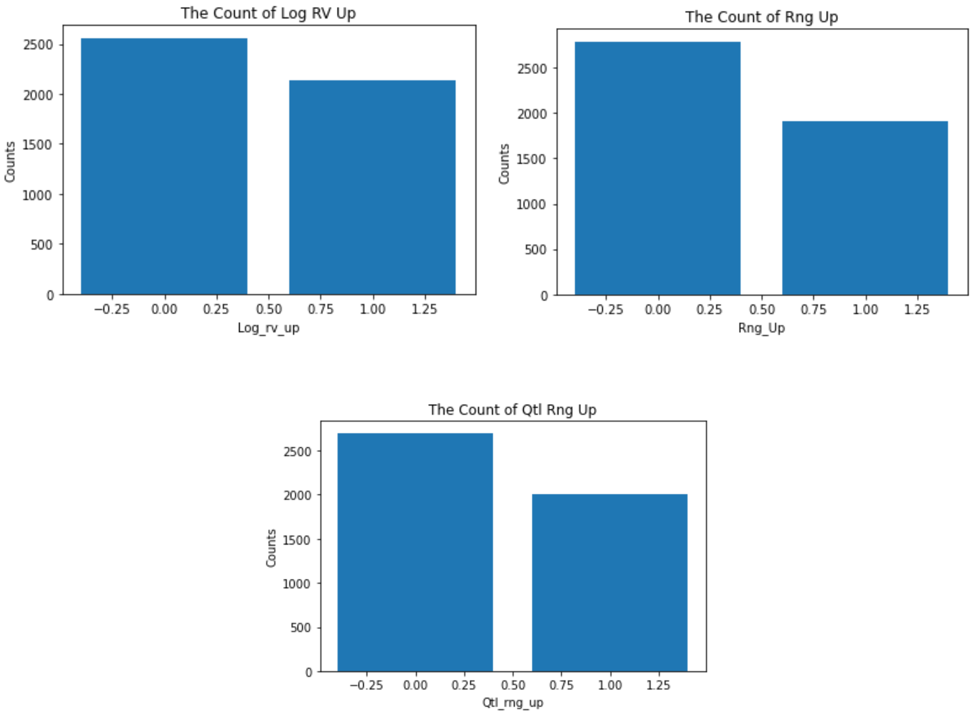
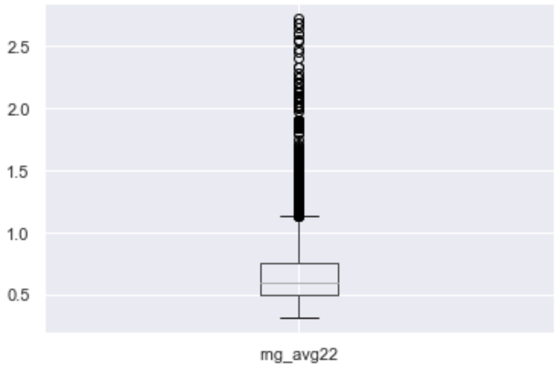


*figure 6 Barcharts for Categorical Variables*

### *Outliers Cleaning*



Box Plot for Average Range over Previous 22 Trading Days *figure 7 Box plot for rng_avg22*

Boxplot is essential to inspect the central tendency of a group of data and figure out the outliers to improve data quality. rng_avg22 is taken as a sample data to process, and the data points that exceed max value (estimate as 1.2) are regarded as outliers here, then these outliers are removed from the entire datasets to improve data accuracy.

*K-means Clustering*

To ensure the members of training and test sets to be selected represent the entire dataset, k-means clustering is adopted (Andrada et al., 2015). The k-means clustering is an unsupervised statistical method used to assign groups according to specific properties that the elements have in common (Andrada et al., 2015).

Using rng_avg_22 as an example, the clustering process started by making the Elbow plot to figure out the efficient number of clusters and avoid over-categorisation or under-categorisation. The horizontal axis of the Elbow plot is the number of clusters, and the vertical axis is a score that can be calculated by measuring the distance between each data point and its closest centroid.

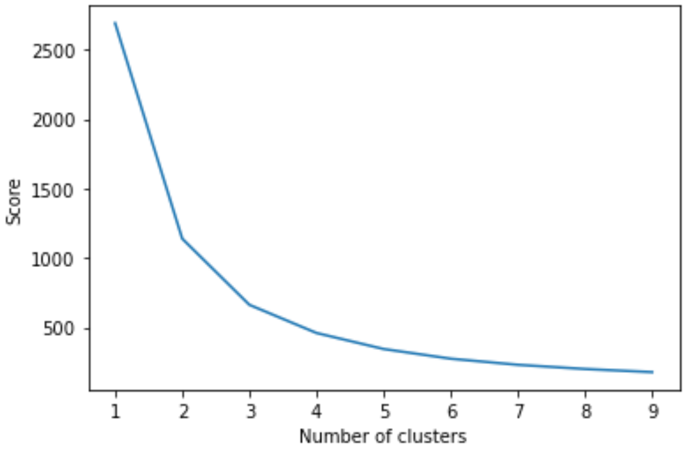


*figure 8 Elbow plot*

As shown in Graph x, when the number of clusters is 3, the score diminishes rapidly and then become almost constant at score 300, this indicates that the increase of the number of clusters will impose inapparent improvements to cluster quality. Thus, based on this insight, the 3 clusters are classified as below:

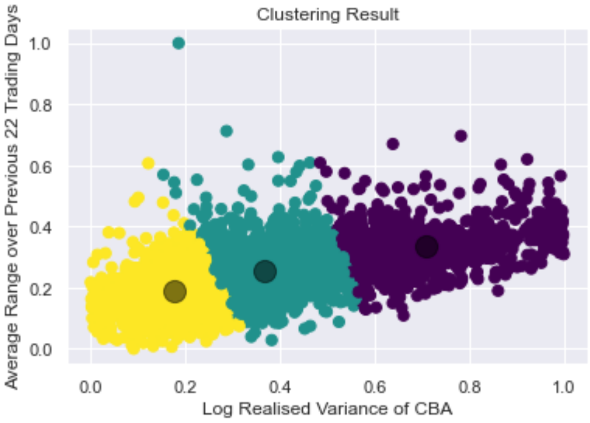


*figure 9 Clusters*

According to this graph, three clusters are displayed, the characters of these clusters are speculated as low return area (yellow), medium return area (green) and high return area (purple).

*Feature Selection and Feature Engineering*

The choice of factors influences the ability of a factor model to predict the prices of log_rv; in particular, the model will be designed based on these three features due to their relatively high correlation with log_rv (Tanskanen, 2018).

| Top 5 correlation with log_rv | |
|---|---|
| log_rv_avg5 | 0.761547 |
| qtl_rng_avg5 | 0.730357 |
| log_rv_avg22 | 0.720919 |

Furthermore, as the scatter plot of qtl_rng_avg_5 shows a polynomial regression curve, so we employ the log-transformation to qtl_rng_avg5 to obtain a more homogeneous variance of this series or to make its distribution more normal, further achieving the linear regression with log_rv for more intuitive modelling in following parts (Lütkepohl & Xu, 2012).

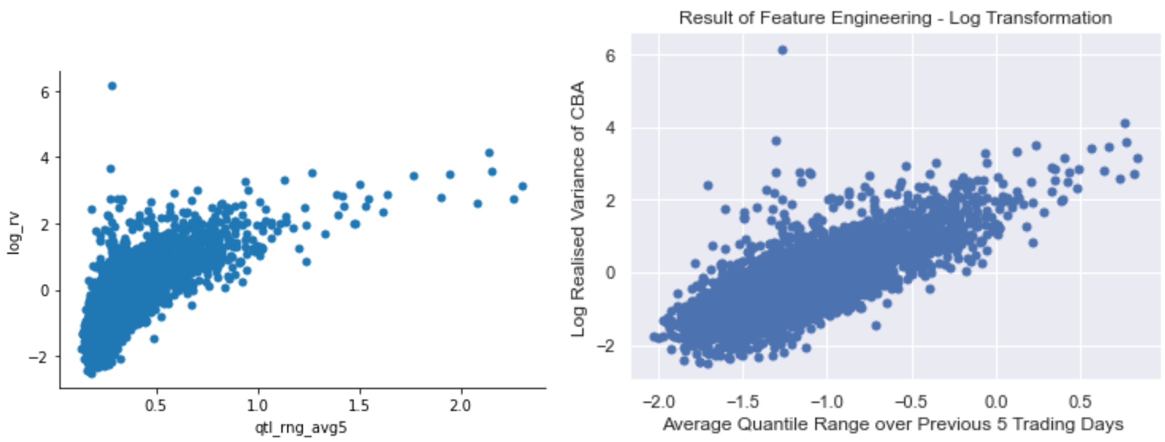


*figure 10 Feature Engineering Results (Before vs. After)*

## Part 2: Forecasting Model Establishment

*Technical explanation of the model*

According to the result of feature engineering, we select log_rv_avg5, qtl_rng_avg5 and log_rv_avg22  as parameters of our linear regression model since each of these features show the best correlation with log_rv in its own category (i.e. the categories hereby refer to the range feature, the log_rv feature and quantile range feature). To begin with, we establish a new DataFrame where outliers data are completely eliminated, and it means we will have data of 4,311 valid trading days to apply for modelling. To construct the model through Python, we import LinearRegression from the Scikit Learn library. After choosing parameters for the y array and x array of the regression model, we perform the training-validation-test split procedure through estimating model on the training set, predicting the observations in the validation set replying on predict function from Scikit Learning library, selecting the model on the validation set and evaluating the performance of the model against the benchmark model. To check if the model meets the basic linear regression assumptions, we adopt fitted value vs residuals plot to execute residual diagnostic, and we will conclude if there is heteroscedasticity by observing the variance and linearity of the variable distribution.

*Model Evaluation*

Based on the clustering result, we assign 30% of the dataset to the testing set and the rest 70% of the dataset to the training & validation set. We then further split ⅓ of the training & validation set to validation and the remaining for the training set. To test the degree of compatibility of our predicted model with the actual dataset, Mean Squared Error (MSE) is employed. MSE can measure how close the predicted regression line is to the actual data points on average, and this distance between each data point to the predicted regression line is called error. MSE is only positive as the errors are squared in the calculation. For our train and validation set, the MSE values are both about 0.3; these results indicate that our regression model is almost desirable without overfitting or underfitting.

We choose MSE as an interpreter for the following reasons. MSE is highly interpretable and understandable for non-professional readers due to its favourable result and is also an excellent metric in the context of optimization because of its properties of convexity, symmetry, and differentiability (Zhou & Bovik, 2009).

*Assumptions and Shortcomings of the Model*

Our model is based on three basic linear regression assumptions

*Linearity*: Pearson's correlation and the respective scatter plots shows a strong relationship between the predictors and the response variables. In the case of the qtl_rng_avg5 predictor, we have applied the necessary log transformation to achieve a strong linear relation.

*Constant Variance*: As seen on the time series plot, the data shows a stationary trend along with the mean of the variable. This shows that the data is Homoscedasticity in nature. The graphs Residual Diagnostic also help examine assumptions of linear regression models (Alvarez 2018). The below Fitted values vs Residuals based on our model illustrates that the Residual does not dramatically change along with the increasing of the fitted value.

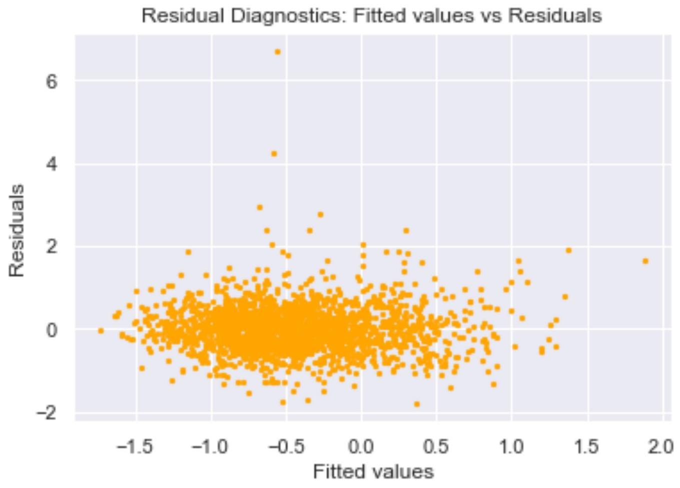


*figure 11 Residual Diagnostics: Fitted values vs Residuals*

*Multicollinearity*: No doubt, the predictors are multicollinear due to lag variables. But we have neglected this violation of linear regression assumption as we are only interested in the final prediction accuracy (Kutner, 2005).

In our opinion, the model could be perfected by incorporating some macroeconomic factors such as interest rates or the overall trend of financial stocks on a particular day. This can only be validated by accumulating more data on these factors, which is readily available in the public domain—finally, modelling and checking the accuracy with these macroeconomic variables—in a nutshell, going back and forth between the modelling and evaluation phase of the CRISP-DM data science process to achieve perfection.

Paradoxically, the benefit of our model is also its most significant shortcoming – The linearity assumption. The linear regression model is excellent at predicting linear relations as long as the data follows a particular path. Still, once it deviates from that linear path, the predictions become inaccurate.

*General Model Interpretation and Example*

In our multivariate linear regression model, the general form of the equation is:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \epsilon$$

and our predicted coefficients are:

| beta 0 | beta 1 | beta 2 | beta 3 |
|---|---|---|---|
| 1.05 | 0.15 | 0.23 | 1.07 |

We can use the capital asset pricing model (CAPM) model to interpret these coefficients as an example. Its general form is:

$$\mathrm{ERi} = Rf + \beta iERm + \beta iRf$$

where: ERi = expected return of investment  Rf = risk-free rate

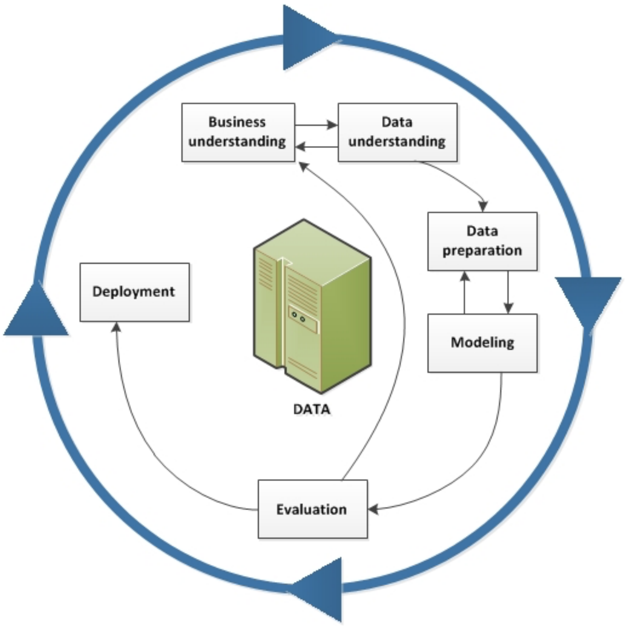
Rm = market risk  Rf = risk - free rate  (Will, 2021)

In this equation, beta as a coefficient indicates the similarity between the investment and the market return. It is an indicator of risk level compared to the market rate, and the positive beta represents a high market risk (Will, 2021). Betas measure the change in the dependent variable (y) given a one-unit change of independent variable (x1, x2, x3) compared to our linear regression model. If beta is large, such as x3, firstly, it forecasts that the increase of x3 price will lead to the positive movement of y; this is also in accordance with the strong and positive correlations between our selected variables and log_rv in EDA. Then, it indicates y is highly sensitive to the change of x3, also reveals the high importance of x3 to the prediction of y. If beta is negative, there will be an adverse movement in y when x increases, and this reminds us the selection of variables in the prediction model is unsuitable.

## Part 3: Reflection

*Work Process and Connections with the CRISP-DM Process*

Cross-Industry Standard Process for Data Mining (CRISP-DM) is an industry-proven guideline of data mining (IBM, n.d.) (Detailed Model Explanations shown in Appendix 2). It acts as both methodology and process model, describing typical phases of a project and making connections between them, further creating a complete data mining life cycle (IBM, n.d.).



*graph 12 Data Lifecycle(IBM, n.d.)*

Overall, the CRISP-DM process was highly effective at guiding us to systematically achieve and evaluate our analysis every step of this assignment. Additionally, the easy-to-understand stages of the entire process benefited us immensely.

*Business Understanding*: Starting with the business understanding phase, before doing any technical data analysis we understood what the Log RV value of a company share means, how it is calculated and did research on what factors might determine the volatility of a share.

*Data Understanding*: Having an initial understanding of Log RV we found it a lot easier to understand the given dataset, what the features were entailing, how useful they might be. Further, at this stage, we were brainstorming about what different models might be useful in predicting future Log RV value.

*Data Preparation*: Next, we prepared our final dataset by incorporating required variables from three different feature sets. Having some inclination of models made us prepare our data according to our chosen models.

*Modelling*: This stage allowed us to evaluate our KNN and Linear regression model, and choose the best model for predicting the Log RV value. At this point, we were going back and forth between the data preparation stage, and making required changes to the dataset for KNN and least-squares algorithm to achieve optimal results.

*Evaluation*: Finally, we evaluated the whole analysis based on the requirement of the business task. This stage allowed us to go back to the first stage and reevaluate our whole process, before submitting our work.

### *Reflections - CRISP-DM Model*

Although CRISP-DM efficiently guides us to complete the data mining, there are still some barriers in our actual working process. Specifically, the business understanding phase can be highly distinct since CRISP-DM is used in different domains (Schröer et al., 2021). The understanding of customers'' business goals and purpose are essential to improve data analysis quality (Schröer et al., 2021). . For example, when we conduct the business understanding, as we lack information about our readers, it is difficult to evaluate whether our data processing and conclusions meet their expectations in contents.

Furthermore, there are three purposes of data science projects, goal-directed, exploratory and data management (Martinez-Plumed et al., 2021). CRISP-DM fits well in goal-directed and process-driven projects, however, when the process becomes more exploratory, the paths will be more varied, and the CRISP-DM model will be insufficient for this flexibility (Martinez-Plumed et al., 2021). Relating to our project, when we operated on data preparation, we found our dataset is combined with categorical and numerical data types. From our perspective, the data classification process is compulsory to add to the CRISP-DM model.

Finally, the popularity of Circular Economy (CE) concepts accelerate the adoption of data recycling (Kristoffersen, 2019).  CRISP-DM is a closed-cycle process without connections with other fields (Kristoffersen, 2019).  It should improve on the restructuring of data understanding and preparation to explore the potential of the dataset and align with overall business and CE goals (Kristoffersen, 2019).

*Reflection - Analytic capability exercise*

When practising the CRISP-DM process model during this project, we have exercised different analytical capabilities from previously learnt knowledge.

In the stage of continuous variable analysis, we performed correlational analysis skills by constructing corresponding graphs to help with our variable selection and data processing. For instance, we adopt heatmaps to examine if there exists multicollinearity between different features, scatter plots to determine linearities between features and log_rv, and the boxplot to clean outliers of our datasets. Further in this stage, we apply skills of data grouping by k-means clustering learnt in Week 7 to classify our cleaned data into three different groups, and the k-means algorithm helps determine the number and types of features which are the most representative.

In the stage of feature engineering, we exercised natural log transformation learnt in Week 7 to stabilise variances of our selected predictors. Such a procedure helps us to initialise the replacement parameter with a well-improved correlation to establish a more reliable prediction model.

In the stage of forecast model establishment, we apply the concept of training, validation and test split learnt in Week 8 to train our model in the training data set, select our model in the validation data set and evaluate the model in the testing set.

In the stage of model evaluation, we executed skills of accuracy measurement by applying residual diagnostics and determination of Mean Squared Error (MSE) of our model. The residual diagnostics graph of fitting vs residuals was selected as it was described as an ideal indicator for us to check variance-consistency and linearity of residuals (Alvarez 2018). While the residual diagnostics can hardly determine the existence of overfitting, we apply

MSE to determine how close the predicted regression line is to the actual value of our datasets.

## References

Alvarez, R. (2018). *Creating Diagnostic Plots in Python.* Robert Alvarez. https://robert-alvarez.github.io/2018-06-04-diagnostic_plots/

Andrada, M. F., Vega-Hissi, E. G., Estrada, M. R., & Garro Martinez, J. C. (2015). Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors. *Chemometrics and Intelligent Laboratory Systems, 143*, 122–129. https://doi.org/10.1016/j.chemolab.2015.03.001

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts.

IBM. (n.d.). *CRISP-DM Help Overview*. https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

Kristoffersen, E., Aremu, O. O., Blomsma, F., Mikalef, P., & Li, J. (2019). Exploring the Relationship Between Data Science and Circular Economy: An Enhanced CRISP-DM Process Model. In *Digital Transformation for a Sustainable Society in the 21st Century* (pp. 177–189). Springer International Publishing. https://doi.org/10.1007/978-3-030-29374-1_15

Kutner, M. H. (2005). *Applied Linear Statistical Models* (5th ed.). Boston: McGraw-Hill Irwin.

Lütkepohl, H., & Xu, F. (2012). The role of the log transformation in forecasting economic variables. *Empirical Economics, 42*(3), 619–638. https://doi.org/10.1007/s00181-010-0440-1

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Tanskanen, A. J., Lukkarinen, J., & Vatanen, K. (2018). Random selection of factors preserves the correlation structure in a linear factor model to a high degree. *PLoS ONE*, *13*(12), e0206551. https://link.gale.com/apps/doc/A566646340/AONE?u=usyd&sid=bookmark-AONE&xid=7eb6c71c

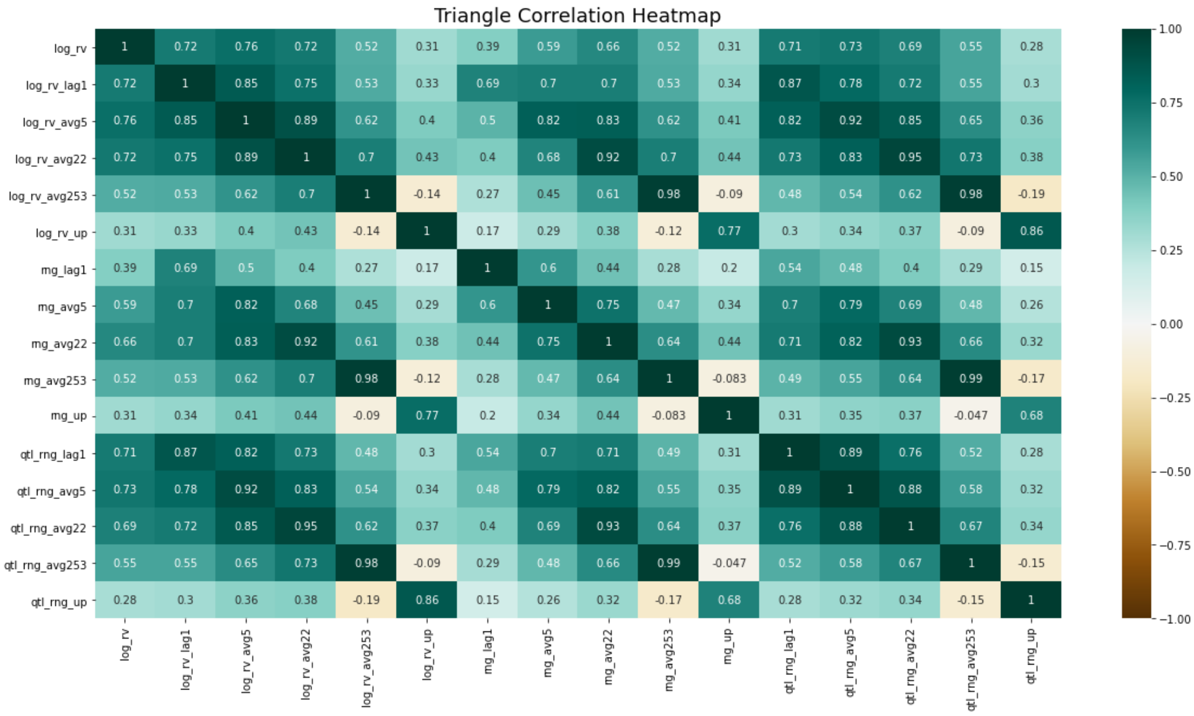
Webster, A. (2013). *Introductory regression analysis: with computer application for business and economics.*Routledge.

Will, K. (2021). *Capital Asset Pricing Model (CAPM).* Investopedia. https://www.investopedia.com/terms/c/capm.asp

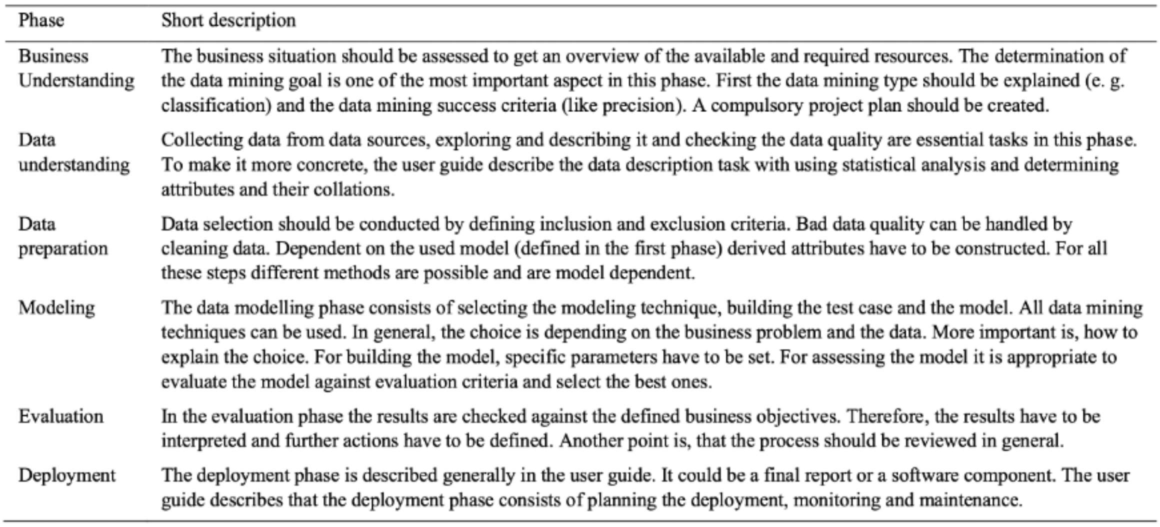
Zhou Wang, & Bovik, A. (2009). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine, 26*(1), 98–117. https://doi.org/10.1109/MSP.2008.930649

# Appendices

## Appendix 1: Correlation Heatmap



## Appendix 2: CRISP-DM process model descriptions (Schröer et al., 2021)



| Phase | Short description |
|---|---|
| Business Understanding | The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created. |
| Data understanding | Collecting data from data sources, exploring and describing it and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describe the data description task with using statistical analysis and determining attributes and their collations. |
| Data preparation | Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent. |
| Modeling | The data modelling phase consists of selecting the modeling technique, building the test case and the model. All data mining techniques can be used. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones. |
| Evaluation | In the evaluation phase the results are checked against the defined business objectives. Therefore, the results have to be interpreted and further actions have to be defined. Another point is, that the process should be reviewed in general. |
| Deployment | The deployment phase is described generally in the user guide. It could be a final report or a software component. The user guide describes that the deployment phase consists of planning the deployment, monitoring and maintenance. |

## Appendix 3:  Data Dictionary

| | |
|---|---|
| log_rv | log realised variance (RV) of CBA |

| | |
|---|---|
| log_rv_lag1 | log RV of the previous trading day [t-1] |
| log_rv_avg5 | average log RV over previous 5 trading days [t-1, t-2, ..., t-5] |
| log_rv_avg22 | average log RV over previous 22 trading days [t-1, t-2, ..., t-22] |
| log_rv_avg253 | average log RV over previous 253 trading days [t-1, t-2, ..., t-253] |
| log_rv_up | binary variable taking value 1 if log_rv_avg22 > log_rv_avg253, and 0 otherwise [t-1, t-2, ..., t-253] |
| rng_lag1 | range of the previous trading day [t-1] |
| rng_avg5 | average range over previous 5 trading days [t-1, t-2, ..., t-5] |
| rng_avg22 | average range over previous 22 trading days [t-1, t-2, ..., t-22] |
| rng_avg253 | average range over previous 253 trading days [t-1, t-2, ..., t-253] |
| rng_up | binary variable taking value 1 if rng_avg22 > rng_avg253, and 0 otherwise [t-1, t-2, ..., t-253] |
| qtl_rng_lag1 | quantile range of the previous trading day [t-1] |
| qtl_rng_avg5 | average quantile range over previous 5 trading days [t-1, t-2, ..., t-5] |
| qtl_rng_avg22 | average quantile range over previous 22 trading days [t-1, t-2, ..., t-22] |
| qtl_rng_avg253 | average quantile range over previous 253 trading days [t-1, t-2, ..., t-253] |
| qtl_rng_up | binary variable taking value 1 if qtl_rng_avg22 > qtl_rng_avg253, and 0 otherwise [t-1, t-2, ..., t-253] |

THE UNIVERSITY OF SYDNEY

# MEETING MINUTES

**Minutes of meeting for** Group 256

Date: 10.24 Time: 19:00 Location: Zoom Meeting

Chairperson: Jingxuan Liu (Daisy)

Minute-Taker: Chen Jing (Steve)

Document tabled:

Present: Ankur Bedi, Daisy Liu, Steve Jing

Apologies: None

| Agenda Item | Key Points | Action | By Whom | When | Communication Strategy |
|---|---|---|---|---|---|
| 1. Task distribution for this group project. (In this meeting particularly for Question 1)<br>2. Scheduling for the next meeting. | Based on the structure of this project, each of the team members will take one part of the following:<br>Q1: Predictive features.<br>Q2: Model building.<br>Q3: Model evaluation & reflection. | We will have at least our very first attempt on Question 1 to find out the predictive features of log RV of CBA. Teams will discuss the answer for Q1 in the next meeting. | Ankur Bedi for Q1 | The next meeting is scheduled at 1 pm on 27th Oct, 2021.<br>Answer for Q1 is expected to be delivered by the next meeting. | The next meeting will be held through Zoom. |

THE UNIVERSITY OF SYDNEY

# MEETING MINUTES

**Minutes of meeting for** _Group 256________________________

Date: 10.27  Time:13:00 Location: __Zoom Meeting________________

Chairperson: __Jingxuan Liu (Daisy)________________________________

Minute-Taker: ____Chen Jing (Steve)________________________________

Document tabled: ________________________________________________

Present: _____Ankur Bedi, Daisy Liu, Steve Jing____________________________

Apologies: ______________________________________________________

| Agenda Item | Key Points | Action | By Whom | When | Communication Strategy |
|---|---|---|---|---|---|
| 1. Research and conduct on statistical testing in tutorials<br>2.Scheduling for the next meeting | By the next meeting, teams are expected to accomplish Q1 essay writing and Jupyter notebook codes for Q2.<br>Besides, teams altogether do research on and conduct statistical testing. | 1) Establish the google document for Q1 essay and finish Q1 writing based on currently available codes and feature choices.<br>2) Writing clustering for Q1<br>3) Establish Q2 codes for Linear Regression (Fitting | 1) Ankur and Daisy<br>2) Daisy<br>3) Steve | The next meeting is scheduled at 1 pm on 30th Oct, 2021. | The next meeting will be held through Zoom. |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  | accuracy and interpreting why we use this model) |  |  |  |

# MEETING MINUTES

**Minutes of meeting for** Group 256

Date: 10.30  Time:18:00 Location: Zoom Meeting

Chairperson: Jingxuan Liu (Daisy)

Minute-Taker: Chen Jing (Steve)

Document tabled:

Present: Ankur Bedi, Daisy Liu, Steve Jing

Apologies:

| Agenda Item | Key Points | Action | By Whom | When | Communication Strategy |
|---|---|---|---|---|---|
| 1. Task review & remaining task distribution<br>2.Scheduling for the next meeting | By the next meeting, teams are expected to accomplish Jupyter notebook codes for Q2, paper writing of Q2 and the reflection. | 1) The coding tasks for Q2 and the technical explanations on paper writing.<br>2) Other parts of Q2 questions on paper writing. | 1)Steve<br>2) Daisy & Ankur | The next meeting is scheduled at 1 pm on 1st Nov, 2021. | The next meeting will be held through Zoom. |

Source: TAFE Access Division “Communication for Business”, 2000

THE UNIVERSITY OF SYDNEY

# MEETING MINUTES

**Minutes of meeting for** Group 256

Date: 11.1  Time:19:00 Location: Zoom Meeting

Chairperson: Jingxuan Liu (Daisy)

Minute-Taker: Chen Jing (Steve)

Document tabled:

Present: Ankur Bedi, Daisy Liu, Steve Jing

Apologies:

| Agenda Item | Key Points | Action | By Whom | When | Communication Strategy |
|---|---|---|---|---|---|
| 1. Task review & remaining task distribution<br>2.Scheduling for the next meeting | Distribution of the rest of our Reflection work was discussed.<br>Reflections are divided by: 1)Work Process and Connections with the CRISP-DM Process. 2) Reflection on our CRISP-DM Process | 1) Finalise reflection 1)<br>2) Finalise reflection 2)<br>3) Finalise reflection 3)<br>4) Accomplish reference list. | 1) Ankur<br>2) Daisy<br>3) Steve<br>4) Everyone | Nov 5th 2021 | |

| | | | | | |
|---|---|---|---|---|---|
| | 3) Analytic skills identification during the modelling process.<br>Reference list is expected to be finished together with reflection.<br>Teams are expected to finalise all the work by this Friday. | | | | |

Source: TAFE Access Division “Communication for Business”, 2000