# MODULE 8 UNIT 1

## Activity submission

Learning outcomes:

**LO1:** Recognise the value in reporting insights extracted from data using statistical analyses.

**LO2:** Illustrate key insights extracted from data.

**LO3:** Analyse the results of an analysis in a business context.

**LO4:** Compile a report communicating the results of your business analysis.

### Name: Matlotlo Magasa

#### 1. Instructions and guidelines (Read carefully)

##### Instructions

1. Insert your name and surname in the space provided above, as well as in the **file name.** Save the file as: **First name Surname M8U1 Activity Submission** – **e.g. Lilly Smith M8U1 Activity Submission.** **NB:** *Please ensure that you use the name that appears in your student profile on the Online Campus.*

2. Write all your answers in this document. There is an instruction that says, “Start writing here” under each question. Please type your answer there. Where there is the instruction “Paste your Tableau visualisation here”, insert the exported image there.

3. Submit your assignment in **Microsoft Word only**. No other file types will be accepted.

4. You will be required to include visualisations that you have made in Tableau in this activity. To do this, export them as an image file and paste them into this document. Tableau has a page that details how to export your Tableau view as an image file.

5. Do **not delete the plagiarism declaration** or the **assignment instructions and guidelines**. They must remain in your assignment when you submit.

**PLEASE NOTE:** **Plagiarism cases will be investigated in line with the Terms and Conditions for Students.**

**IMPORTANT NOTICE:** Please ensure that you have checked your course calendar for the due date for this assignment.

##### Guidelines

1. There are 15 pages and 7 questions in this assignment.

2. This activity submission relies heavily on work done in the IDE activity in the previous unit. If you have not done so already, return to the IDE notebook and complete this activity before attempting to answer the questions in this assignment.

3. You will be required to download a data file, titled HOUSING.csv, in order to interact with the data in Tableau. Return to the Module 8 downloads folder and save a copy of the data file before attempting to answer the questions.

4. Make sure that you have carefully read and fully understood the questions before answering them. Answer the questions fully, but concisely, and as directly as possible. Follow all specific instructions for individual questions (e.g. “list”, “in point form”).

5. Answer all questions in your own words. Do not copy any text from the notes, readings, or other sources. **The assignment must be your own work only.**

|  |
| --- |
| **Plagiarism declaration:** |
| **1. I know that plagiarism is wrong. Plagiarism is to use another’s work and pretend that it is one’s own.**  **2. This assignment is my own work.**  **3. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.**  **4. I acknowledge that copying someone else’s assignment (or part of it) is wrong and declare that my assignments are my own work.** |

#### 2. Mark allocation

Each question receives a mark allocation. However, you will only receive a final percentage mark and will not be given individual marks for each question. The mark allocation is there to show you the weighting and length of each question.

Introduction 10

Question 1 20

Question 2 20

Question 3 20

Question 4 20

Question 5 20

Conclusion 10

**TOTAL 120**

#### 3. Questions

Note:

Review the requirement for each question then use the report template in Section 4 to provide your answers.

After presenting your initial findings to the Southern Realty management team, your line manager has asked you to provide your interpretation of the results for each analysis, which will serve as a final report to be distributed among the board of directors. You have specifically been asked to address the business questions posed at the first weekly meeting and suggest how the results might be used to make data-driven decisions for the company’s benefit.

You will need to draft a single report with seven distinct subdivisions, which provide an introduction to your analyses, a conclusion on your findings, and detailed responses to each of the business questions:

1. **Introduction:** The opening paragraph should outline why these analyses were performed and the business value that can be gained from performing the recommended statistical tests.
2. **Question 1:** Report on the findings of the factor analysis.
3. **Question 2:** Report on the findings of the cluster analysis.
4. **Question 3:** Report on the findings of the multiple linear regression.
5. **Question 4:** Report on the findings of the logistic regression.
6. **Question 5:** Report on the findings of the classification tree analysis.
7. **Conclusion:** The conclusion should provide a bulleted list that summarises some of the main findings and what data-driven decisions the business might make to ensure the profitability of the investment opportunity.

The following subsections describe the specific questions that the Southern Realty management team would like answered for their presentation to the board of directors. In instances where Tableau visualisations need to be created, embed these visualisations in the report, and include an appropriate introduction and caption for each. For each question, complete the report in concise, coherent paragraphs and only use bulleted lists where requested to do so.

##### Question 1: Factor analysis

One of the first questions the management team posed was whether there are any underlying relationships between the variables in the data set, and whether these could be captured by latent, conceptual factors. To answer this question, you performed a factor analysis. Now you need to provide your interpretation (refer to the casebook to help you address this question). The management team specifically wants you to consider the following points:

* The elbow rule and Kaiser’s criterion – which were used to select the number of factors to include in the final model – suggested different outcomes. Where Kaiser’s criterion suggested three factors for the final model, the elbow rule suggested five. Provide reasoning as to why only three factors were included in the factor model. In your response, pay specific attention to the amount of variance that is explained.
* Although the majority of the variables loaded highly onto a single factor in the unrotated factor model, you argued that factor rotation would still simplify the final interpretation of the model. Provide two examples of where factor rotation improved the interpretability of the results and explain why this is the case.
* While the majority of the variables had high communality values – suggesting that these variables are adequately explained by the factors – other variables showed significantly lower values. What could these low communality values indicate?
* Lastly, based on the final rotated factor model, the team has requested that you name the factors based on the variables that load highly onto each. In your description, create a bulleted list and substantiate why you believe it makes sense, conceptually, to group the respective variables under each factor you name.

##### Question 2: Cluster analysis

Next, you moved on to determining whether it was possible to group observations in the data set into defined clusters that share similar characteristics. This would provide the company with a high-level view of the types of properties that exist in the San Francisco Bay Area and highlight potential areas that could prove profitable investments by appealing to different customer incomes.

The clustering method you found the most effective was the non-hierarchical *k*-means clustering method, the results of which are provided in the casebook. The team has asked you to do the following:

* Profile each cluster, referencing the average values of each variable in that cluster.
* Note any clusters describing areas that might accommodate higher-income individuals, as this is the target market for Southern Realty. In addition, mention any variables – other than owner income – that Southern Realty should consider in a property to ensure that they focus on properties that would interest the company’s target market.

##### Question 3: Multiple linear regression

Next, Southern Realty is interested in determining the variables that had the most influential impact on determining the median house value. The team has hypothesised that there is likely an association between the value of the property and the income of the property owner. In other words, individuals with a higher income would likely invest in a more expensive property. To answer this question, you have decided to visualise this relationship in Tableau by means of a scatter plot and trend line.

To do this, drag the “Median income” measure into the “Columns” field and the “House value” measure into the “Rows” field. Remember to change the variables from “Measure” to “Dimension” to visualise individual data points on the scatter plot. Add a trend line to the graph by right-clicking on the plot, selecting “Trend lines”, and clicking on “Show trend lines”. Rename the worksheet “Median income versus house value”.

Based on this visualisation, it is obvious that some form of relationship exists between median income and house value. To better define this relationship, you performed a simple linear regression analysis with median house value defined as the dependent variable, and median income defined as the independent variable. However, you were also interested in understanding whether any other independent variables significantly impacted the median house value. Therefore, you also generated a multiple linear regression model that included all significant variables within the model. Refer to the casebook for the results of these analyses.

The team has now asked you to provide your interpretation of the results, specifically considering the following:

* Referring to the simple linear regression model that was generated, explain the results of the analysis and why it was deemed necessary to include additional independent variables in the model.
* What value can the business derive from ranking the variables in the multiple linear regression model, in order of importance? Substantiate your answer by providing the ranked variables and explaining how Southern Realty can gain value from them.
* Discuss whether the final model violates any inherent assumptions of linear modelling. Ensure that you explain any conclusions by referring to the analysis.

##### Question 4: Logistic regression

While the median house value provides an indication of the property value in a specific area of San Francisco, it does not provide an indication of whether the house will sell. As a result, Southern Realty has asked you to identify the variables that can help predict this outcome. Having received data on house sales (with a categorical outcome of either “Yes” or “No”), you decided to perform a logistic regression analysis and create some informative visualisations. Refer to the casebook for the results of this analysis.

First, the company is attempting to decide whether to invest in selling newer or older properties, based on a property’s distance to the ocean. To explore how house sales differ based on property age and proximity to the ocean, you decide to plot a bar graph in Tableau.

To do this, drag the “Ocean proximity” dimension into the “Columns” field, and drag the “House sale” dimension and the “Housing median age” measure into the “Rows” field. Remember to change the “Housing median age” from SUM to Average. Change the colour of the graph by dragging the “House sale” dimension over the “Color” mark in the “Marks” pane. Rename the worksheet “House sales by property age and proximity”.

The management team has asked you to provide a summary of the data visualisation and logistic regression results, and to explain how to interpret these from an investment perspective. You have specifically been asked to consider the following:

* Analyse the Tableau visualisation and indicate how house median age and ocean proximity relate to whether a house is sold.
* What can you infer from the final logistic regression model, considering the nature of the relationship between the significant independent variables and whether a house was sold? If you wanted to rank the relative importance of the variables in the analysis, how might you go about this?
* Discuss the fit of the model by referring to the null and residual deviances, as well as the predictive accuracy of the model.

##### Question 5: Classification tree analysis

Finally, the team is interested in knowing whether it is possible to draft a series of guidelines (or rules) that can help to predict whether a house will sell based on the set of independent variables in the data set. To answer this question, you decided to create a classification tree model, the results of which can be found in the casebook.

The team wants you to provide insight into what the data means and whether the model you have created is useful in classifying whether or not a house will sell. You have been asked to do the following:

* Provide a bulleted list of decision rules that can be made when attempting to classify whether or not a house will sell.
* Discuss whether the model you created generalised well to unseen data and if it could be useful in classifying observations from data that was not used to train and create it. How might you further test whether the model is a good fit?

#### 4. Final report

Complete the following sections by answering the questions posed for each type of analysis.

##### Introduction

(Max. 250 words)

Start writing here:

Southern Realty is seeking to expand its real estate company to the USA in the San Francisco Bay area. Before such an investment is made, the company needs to draw insights into the market in this region in order to assess which areas to invest in as well as which market variable to focus on in order to maximise on profit.

A sample of data was used based on real estate in the area to do business analysis of Southern Realty. Various business analytic tools were configured to help answer and provide insights for the company.

A factor analysis was performed to simplify and summarise the dataset in order to investigate whether there are any relationships between variables that may be better explained through factors. A cluster analysis was performed to gain insight of the similar characteristics of the properties types and how they appeal to the Southern Realty target clients. In order for the company to make an informed decision as to what property attributes to invest more in to increase profits from the housing value, a multiple regression model was computed. Logistic regression was used as means to predict whether a house sold in the past year using different variables. The importance of this being to establish the variables that maximise on house sales. To farther asses the profitability of investments in the different property, guidelines were established that would help predict whether a house would sell or not.

The results of the analysis are provided in this report.

##### Question 1: Factor analysis

(Max. 600 words)

Start writing here:

It is important to condense the variables of similar characteristics into factors in order to understand the underlying relationships in these and how they may influence the success of the company’s expansion in the USA. Variables with high correlation between each other have similar characteristics and are likely to be part of a single cluster or factor. The factors will allow the company to better understand the factors that drive sales within the real estate environment in San Francisco.

From the correlation matrix, the highest correlation between variables is the Households & Total Rooms. The correlation between these variables are indicative of the variables that may be grouped into factors.

To determine the number of factors that will describe the dataset, a factor model was computed using Singular Value Decomposition to determine the components and their respective eigenvalues. Two methods were used to determine the number of factors that will be suitable to explain the highest variance within the data set. These are the Kaiser criterion and the Scree plot.

The Kaiser analysis determines the factors of the data set by selecting the number of components with eigenvalues equal or greater to 1. According to this criterion there are three factors that will describe the dataset. These 3 factors explain a total variation of 85.2% within the dataset.

In the second method, the eigenvalues are plotted on a scree plot. The curve plateaus at 5 factors, which is the suggested number of factors for the dataset. Although five factors will explain 98% of the variation in the dataset, five factors will make interpretation of these factors difficult to interpret. Therefore 3 factors explaining 85.2% of variation is a better model to consider and compromise over a model which explains more variation but does not make practical sense for the company to use.

The estimation of factor loadings for each independent variable is necessary to determine the factor to which it is to be placed. Latitude had a distinctively strong association to factor 1. Other variables loaded highly in one of the factors as well, however, the Median\_Income displayed an applicable, however weak loading on the factor 1. House\_Median\_age displayed a weak cross-loading between factors 0 and 1. This means that the House\_Median\_age does not load high enough (significantly above 0,5) on an individual factor. A weak association to a single factor may cause confusion on interpreting the model. To remedy this, a Varimax factor rotation was performed to improve and increase an association to a factor for each of these variables. This was done by plotting the variables onto a Cartesian plane and orthogonally rotating the axes until the variables load highly on a particular factor.

The factor rotation improved the factor loadings for Median\_Income, but only by an additional loading of 0.02. The House\_Median\_age factor loading did not improve after rotation.

The communalities for the different variables were also determined. Again, it’s the House\_Median\_age and Median\_Income variables with low communalities while the other variables have significantly high communality (closer to 1.00). Low communalities of the variables, i.e. communality less than 0.5, indicate that there is a low variance across the factors for these variables, meaning that the variable spreads almost evenly across all or some factors. This makes it difficult to distinguish between which factor a variable may associate with. Low variation of the variables will create confusion in interpretation, and therefore these variables with may be removed from the model. Thus House\_Median\_age is recommended to be removed from this model.

The factors were named as follows based on the similarity of the variables of these factors:

Table 1: Business Factors



##### Question 2: Cluster analysis

(Max. 800 words)

Start writing here:

Southern Realty seeks methods to maximize profits in the housing venture that they will be investing in. To do this a cluster analysis was done on the data set to determine any clusters that may give insight into where the business may draw maximum profits from as well as establish their target market preferences.

Two clustering methodologies were used, these are namely hierarchal methods and the k-means algorithm. Hierarchal methods measure dissimilarity of objects through distance metrics.

Hierarchal Algorithm Performance:

The single linkage method was the worst performing of the three hierarchal methods. The dendrograms produced unclear and compressed tree branches and proved poor in distributing different observations in each cluster. Manhattan was the distance metric dendrogram suggested four clusters and provided a better visual representation of the dendrogram over the Euclidean and correlation measurements. The correlation metric was the worst of the metrics due to its undiscernible dendrogram branches as well as only suggesting 2 clusters from the dendrogram.

The Complete linkage method and the Centroid method were the next tried algorithm in exploring the clustering of the dataset. The dendrograms of these two algorithms suggested that there should be 4 clusters. Complete linkage and Centroid algorithms were improved methods from the single linkage, showing better discerning and improved dendrograms. However, the amount of distribution of the observations across clusters is not satisfactory in either of these methods.

The non-hierarchal k-means algorithm proved to be the best method over the hierarchal methods. The algorithm iteration produced 4 clusters with a relatively even distribution across the clusters. Cluster 0 is seen to have the highest frequency of observations while cluster 1 has the least.

A standardized cluster means graph was plotted to observe whether there is any variation of variable means in each cluster. Clustering seeks to decrease the variation with observation in each cluster while maximizing the variation between different clusters. The ‘Total Bedrooms’ and the ‘Population’, ‘Median income’ and the ‘Household’ appear to have the greatest mean variation in the different clusters. The House Median Age has little variation in mean for clusters 0 and cluster 2.

An ANOVA test was performed to further investigate the statistical significances of the cluster means. The F-static was explored with the p-value. A high F-statistic and a p-value less than 0.1 denotes a significant variable within the clusters. The longitude, latitude and median income displayed significantly low F-static values with p-values far above 0.1. These variables are thus non-significant to the clustering model. The significant models are the total rooms, total bedrooms, population, households and the housing median age. It is important to note that the median age displayed a lower F-statistic than the other significant variables. This is likely due to the low variation in the cluster means of cluster 0 and cluster 2. This means that the median age of the houses contributes less in differentiating between these two clusters. The Total Rooms cluster-means differs significantly across the clusters (i.e. has the highest cluster mean variation across clusters) as its p-values is closest to 0 and has the highest F-statistic.

The clusters can be labelled as follows:

Cluster 0: Small to medium sized houses with small to medium community/neighbourhood

Cluster 1: Large sized houses with large community

Cluster 2: Small-sized houses with small community

Cluster 3: Medium to large sized houses with medium to large community

The target market for Southern Realty are higher income clients. Should the company want to appeal to these clients, they should appeal to these clients. The higher income clients are distributed between cluster 1 and cluster 3, with cluster 1 with the higher income clients. These clusters are synonymous with larger houses (higher number of total rooms and bedrooms) as well as a denser population. The ‘dense’ population may due to a larger block or larger neighbourhood. The company should also invest in relatively newer to moderately aged houses in order to attract their target market.

##### Question 3: Multiple linear regression

(Max. 700 words)

Start writing here:

A regression model was necessary to determine the most influence on the median house value on the block. This will give insight into which variables will have greater influence on housing value. A correlation matrix was computed to determine the variable with the strongest relationship with the median house value. The clients’ median income had the most correlation with the median house value at 0.697 (69.7% correlation).

A simple linear regression model was plotted between median house value and income. Figure 1 shows this relationship.

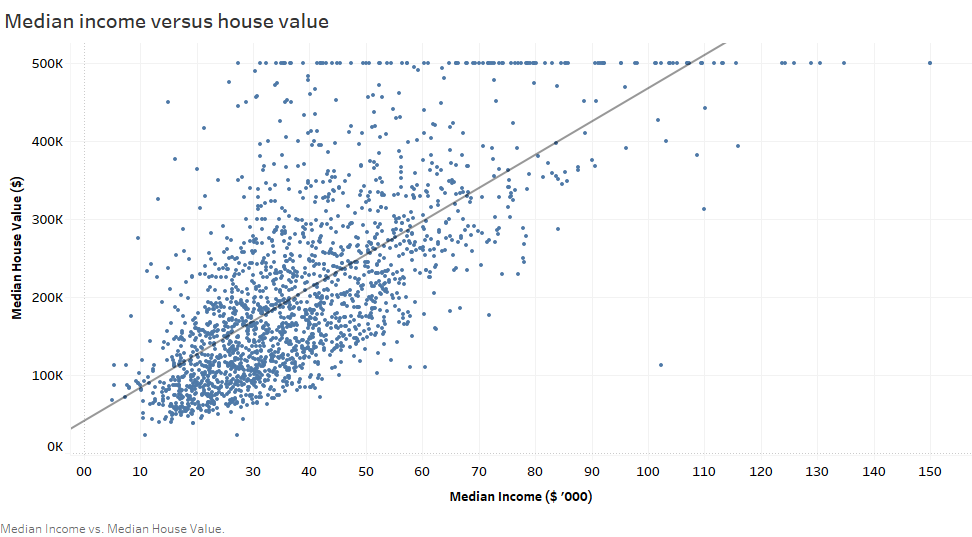


Figure 1: Median income versus house value

The plot shows a positive relationship between the median income and the median house value. This is intuitive as we expect the higher earning clients to buy higher priced houses relative to the lower earning clients. The R2 value of the plot is 0.486, and therefore the median income explains 48.6% of the variation in the median value of the households.

The relationship of the income to the house values displays only medium strength. The regression model would be more reliable when this strength in this relationship is a stronger one. It became necessary to farther explore increasing the strength of this model by investigating how other variables may influence the prediction of the house value. An increase in other independent variables will result in an increase in the R2 value.

When adding the rest of the variables, the R2 value increased to 0.670 (67.0%). The full model can collectively predict 18.4% *more* variation in the median house value than the simple regression model. A large R2 is ideal for the model, however, including all independent variables in the model can increase overfitting and have multicollinearity. Multicollinearity between variables impacts the model accuracy and may skew the outcome. The correlation matrix heat map displayed that there are a few independent variables that are correlated with each other and can possibly add to multicollinearity of the model.

The company has indicated a significance level of 10%. A backward selection stepwise regression was performed to eliminate non-significant variables. All variables have a p-value of 0, except for Total Room, rendering it non-significant for the analysis. After the removal of this variable, there was no change or improvement in the R2  value.

The beta coefficients in the regression model allows us to determine the relative importance of each variable in how it influences the median housing value. In order to accurately rank these variables, the independent variables need to first be standardized onto one scale. The variables have not been standardized and therefore cannot be ranked by importance. However, assuming the coefficients were standardized, then they would rank as follows:

Table 2: Multiple Regression-Variable relative importance



Ranking the variables for Southern Reality will give investors a sense of which variables are contribute greatly in the price of the housing. Such insights will allow Southern Reality to accurately price the value of the house bases on whether the house is near the ocean or bay for example.

Assumptions of linear regression:

The residual plots computed show that the homoscedasticity and the independent errors assumption has been violated. In the Residuals vs Fitting plot in Figure 2, The fitted value errors do not show a constant variance centred around the y-axis. This is an indication that multicollinearity indeed exist in the model. This is farther noticed in the individual variable residual plots. A pattern resembling a straight line can also be observed on the plot.

This result is expected as the correlation heat map showed a great amount of multicollinearity between many variables, however only one variable was eliminated through the backward selection stepwise regression. It would be of value that a forward selection stepwise regression be performed as a way of a better method of detecting and eliminating variables that may have multicollinearity.

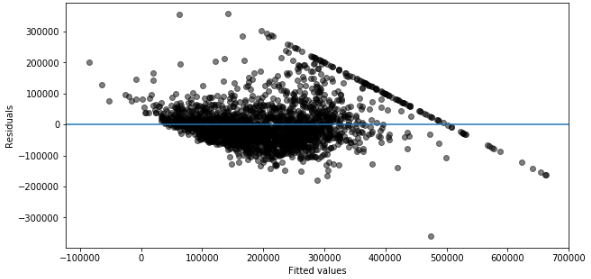


Figure 2: Residual vs Fitting Plot

The histogram has a smooth bell curve which is indicative as of a normal distribution. The data points on the Q-Q Plot fall along a straight line which suggests a normal distribution. The lighter right trail lifted from the line; however, this suggests that the distribution is skewed right. The normality assumption is therefore obeyed.

##### Question 4: Logistic regression

(Max. 700 words)

Start writing here:

Investigating the House sales by property age and Proximity, it can be deduced that many of the older houses do not sell across all ocean proximities. The houses less likely to sell are all over 30 years old. Newer houses (less than 15 years old) sell more when they are located less than an hour away from the ocean as well as when they are inland. It is important to note that there are no new-built houses located near the bay and as a result there is almost a 50-50 chance of these houses selling. Refer to Figure 3 below:

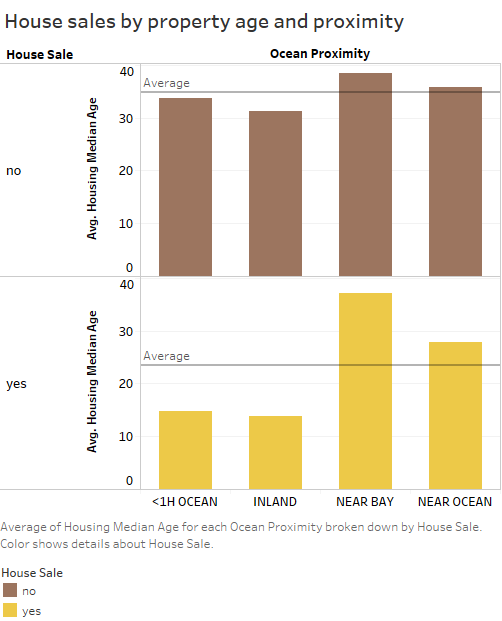


Figure 3: House sales by property age and Proximity

A logistic regression model was used to predict whether houses would sell based on various predictor variables. To better understand the relative importance of the predictor variables, it is important to standardize the variables to one scale then exponentiate the coefficients for easier interpretation. The variables may be ranked once this has been done. Assuming the variables have been standardized:

Table 3: Exponentiated Logistic coefficients



The residual deviance (-2LL) is a means of testing how well the maximum likelihood estimation fits the model, the lower the -2x(Log Likelihood), the better the model fit. The null model of the logistic regression has a deviance of 2754.13 and upon adding all independent variables, the residual deviance of the ‘Full Model’ improved to 1392.2. The final model deviance was recorded to be 1396.6, this is a 49.3% reduction from the null model deviance. However, this improvement is still significantly higher than a perfect model of 0 deviance.

To farther investigate the predictive accuracy of the final model, the Hosmer-Lemeshow chi-squared models were computed for both the full and final model. Low chi-square values imply that the actual predicted values from the model and the observed values are strongly associated with each other and have very little differences between each other. The chi-squared and p-values were 1.0 and 0.0 respectively for both models. This measurement is indicative of a poor model. Based on these results, the feature selection method did not significantly improve the model accuracy on the final model as there was no change in the chi-squared and p-values.

The Full model was calculated to have an accuracy of 85.8% while the final model is 86% accuracy. The accuracy comparison supports the Hosmer-Lemeshow estimates as well as the log-likelihood models in that there has been very little improvement of the final model from the full model.

Upon investigating the different predictive accuracies of the two outcomes (‘yes’/’no’), there’s a relatively small imbalanced misclassification among the classes. The accuracy in predicting no sales is 88. 69% (11.31% misclassification) and an accuracy of 82.74% (17.26% misclassification) for predicting occurring sales. Therefore, the classification of the minority class, i.e. positive sales, is likely to have more predictive error.

The pseudo R2 value has been computed for the final model and is 0.4929. Therefore, the predictor variables in the final model only explain 49.29% of the variation in whether a sale has occurred on not. This is only a moderate strength in predicting the likelihood of the outcomes. Bases in this result the model is not a good one on predicting house sales.

Although there have been no evidence suggesting that removing non-significant independent variables from the model has improved the model, the model can still be improved for better prediction. From the correlation data between independent variables (from computing multiple regression), it can be observed that there is a high correlation of 89.27% between ‘Total Rooms’ and ‘Households’. These two variables are both included as significant predictor variables in the logistic regression model. In the attempt to improve the final model, it will be highly recommended that one of these variables be removed in the final model as logistic regression does not perform well when there is multicollinearity between independent variables. A suggestion would be to remove the variable with the lower relative importance or coefficient.

##### Question 5: Classification tree analysis

(Max. 500 words)

**Marker’s Note: Please note that the total for the test set on the confusion matrix is 505 instead of an expected 500 (i.e. 25% of 2000 observations). I have therefore used this value in my calculation.**

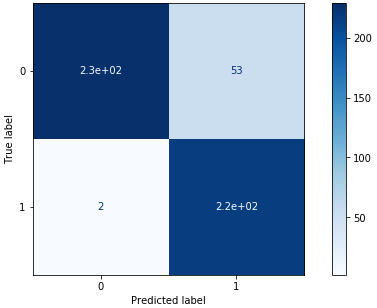
Start writing here:

A classification-tree has been computed to determine the sales outcome of houses in San Francisco. The training error graphs shows that the model begins overfitting around 13 nodes, however, the cv error suggests that for optimal minimal error, 6 nodes can be used for the final model.

The classification tree will enable the sequential explanation as to how a decision to an outcome was reached.

The guideline rules in determining the outcome of house sales are as follows:

1. Should the median age of the houses within a block be 19.5 years old or less, the house will sell.
2. If the median age of the houses is more than 19.5 years old and is near the ocean, the house will sell.
3. If the house median age is more than 19.5 years old, the house *not* near the ocean and *not* near the bay, as well as have total rooms of 3970 or less, the house will not sell.
4. If the house median age is more than 19.5 years old, *not* near the ocean and *not* near the bay, however, have total rooms more than 3970, the house will sell.
5. If the house median age is more than 19.5 years old, *not* near the ocean butnear the bay with a housing median age of 22.5 years or less then the house will not sell.
6. If the house median age is more than 19.5 years old, *not* near the ocean butnear the bay, with a housing median age more than 22.5 years then the house will sell.



The model misclassified the test data with 53 false positives and 2 false negatives. The accuracy of the model to predict the sale (‘Yes’) of a house resulted to be 99.09% ; while that of a lack of sale (‘No’) resulted in an accuracy of 81.27% . This means that there has been more misclassification of the ‘No’ (0) than ‘Yes’ (1).

The model, therefore, is better at predicting whether a house *will* sell more than it is predicting whether a house won’t sell.

To determine whether this model is indicative of a good model the proportional chance criterion is determined.

Proportional chance criterion:

Because the lack of sales or “No’s” have the most observation of 283, this will be used as the reference for the proportional chance criterion-

The model hit rate, 89% is more than the proportional chance criterion. Therefore, pruning the tree model to 6 nodes, the model displays an ability to generalise to unseen data, thus making it a good fit

Thus, the 6-node tree may be used as a method of classification to determine which part of the San Francisco Bay area to invest in.

##### Conclusion

(Max. 250 words)

Start writing here:

In order for Southern Realty to expand its business in the San Francisco Bay area of the USA, a comprehensive business analysis of the market is necessary. Using various analytic models, the San Francisco real estates market was analysed to allow the company to gain insight on how to invests and in which areas to invest in.

The following results were obtained:

* There are three factors that summarize the real estate market in San Francisco, these are namely the property size, the location and the wealth of the potential clients. For the company to maximize on profits, they should focus on these factors that may appeal to their target market
* Cluster analysis provided insight into the different types of property and which appeal to Southern Realty’s target market, which are the high-income clients. To appeal to this market, Southern realty should invest in medium to large sized houses with larger neighbourhood/community
* High-end clients prefer likely to buy high-valued housing. Therefore, housing value will increase when the property is firstly, near the ocean or at least an hour away from the ocean. High valued housing is also found near the bay. The company should invest in such property to maximise on profit.
* To specifically determine which house to invest in in the neighbourhoods mentioned above, the housing median age can be an influence in whether a house will sell. A set of guideline rules have been made to increase the chances that Southern Realty will invest in houses most likely to sell.

s

#### 5. Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No submission** | **Poor** | **Good** | **Excellent** |
| **Introduction**  *The submission adequately outlines why the analyses were performed and addresses the business value that can be gained from performing the recommended statistical tests. The submission is written in a coherent, logical manner.* | No submission or attempt. (0) | The student either outlines why the analyses were performed or addresses the value that can be gained from performing the analyses, but not both.  AND  The submission has some logical structure and is clear enough to comprehend. (5) | The student both outlines why the analyses were performed and addresses the value that can be gained from performing the analyses.  AND  The submission has some logical structure and is clear enough to comprehend. (7) | The student both outlines why the analyses were performed and addresses the value that can be gained from performing the analyses.  AND  The submission is structured exceptionally well in terms of logic and clarity. (10) |
| **Question 1: Number of factors**  *The submission provides insightful reasons for the selection of three factors for the model.* | No submission or attempt. (0) | The reason for choosing three factors is mostly correct. (2.5) | The reason for choosing three factors is correct, but incomplete, or lacks some insight. (3.5) | The reason for choosing three factors is correct and comprehensive and displays substantial insight. (5) |
| **Question 1: Factor rotation**  *The submission explains why factor rotation is helpful in improving the interpretability of the model, using two examples in support of the argument.* | No submission or attempt. (0) | The student makes an adequate argument for the use of factor rotation to improve the interpretability of the model and uses two examples to support their answer. (2.5) | The student makes a good argument for the use of factor rotation to improve the interpretability of the model and uses two examples to support their answer. However, the insight offered is limited or incomplete. (3.5) | The student makes an excellent and insightful argument for the use of factor rotation to improve the interpretability of the model and uses two examples to support their answer. (5) |
| **Question 1: Communality**  *The submission explains what low communality values might indicate about the model.* | No submission or attempt. (0) | The explanation of what low communality values might indicate about the model is mostly correct. (2.5) | The explanation of what low communality values might indicate about the model is correct, but the insight offered is limited or incomplete. (3.5) | The explanation of what low communality values might indicate about the model is correct and offers substantial insight. (5) |
| **Question 1: Naming the factors**  *The submission adequately names and describes the three factors.* | No submission or attempt. (0) | The student names the three factors correctly and provides partially correct descriptions. (2.5) | The student names the three factors correctly and provides correct descriptions, but the conceptual understanding is limited. (3.5) | The student names the three factors correctly and provides correct descriptions, displaying a good conceptual understanding of the underlying structure. (5) |
| **Question 2: Profiling clusters**  *The submission correctly profiles each of the four clusters by interpreting the variables within each and referencing the mean values for each variable.* | No submission or attempt. (0) | The student profiles each of the four clusters by interpreting the variables within each, but does not profile all clusters correctly. (7.5) | The student correctly profiles each of the four clusters by interpreting the variables within each. However, the insight offered is limited. (10.5) | The student correctly profiles each of the four clusters by interpreting the variables within each and delivers substantial insight regarding the characteristics of each cluster. (15) |
| **Question 2: Higher-income clusters**  *The submission correctly identifies whether any clusters would describe housing characteristics that might accommodate wealthier clientele and elaborates on the other variables Southern Realty should consider in a property.* | No submission or attempt. (0) | The student identifies the correct clusters, but does not provide an explanation of the other variables that Southern Realty should consider in a property. (2.5) | The student identifies the correct clusters and provides an explanation of the other variables that Southern Realty should consider in a property. However, this explanation is limited or incomplete. (3.5) | The student identifies the correct clusters and provides a comprehensive explanation of the other variables that Southern Realty should consider in a property. (5) |
| **Question 3: Tableau visualisation**  *The submission provides the correct Tableau visualisation.* | No submission or attempt. (0) | The image supplied contains the incorrect graph type or the incorrect variables are used. (2.5) | The image supplied contains the correct graphs using the correct variables, but does not adhere to all instructions. (3.5) | The image supplied contains the correct graphs using the correct variables and adheres to all instructions. (5) |
| **Question 3: Simple linear regression**  *The submission correctly explains the results of the simple linear regression model and gives accurate reasons for including additional independent variables in the model.* | No submission or attempt. (0) | The student correctly explains the results of the simple linear regression model, but does not give accurate reasons for including additional independent variables in the model. (2.5) | The student correctly explains the results of the simple linear regression model and gives accurate reasons for including additional independent variables in the model. However, the reasons lack insight. (3.5) | The student correctly explains the results of the simple linear regression model, and gives insightful reasons for including additional independent variables in the model. (5) |
| **Question 3: Multiple linear regression**  *The submission accurately ranks the independent variables in order of importance and provides an adequate explanation of the value that can be gained from this.* | No submission or attempt. (0) | The student either accurately ranks the independent variables in order of importance or provides an accurate explanation of the value that can be gained from this, but not both. (2.5) | The student accurately ranks the independent variables in order of importance and provides an accurate explanation of the value that can be gained from this, but the explanation is limited or incomplete. (3.5) | The student accurately ranks the independent variables in order of importance and provides a correct and comprehensive explanation of the value that can be gained from this. (5) |
| **Question 3: Assumptions**  *The submission concludes whether the assumptions of multiple linear regression have been violated.* | No submission or attempt. (0) | The student correctly concludes whether the assumptions have been violated, but only provides limited reasoning. (2.5) | The student correctly concludes whether the assumptions have been violated, but the reasoning lacks insight. (3.5) | The student correctly concludes whether the assumptions have been violated, and provides correct and comprehensive reasoning. (5) |
| **Question 4: Tableau visualisation**  *The submission provides the correct Tableau visualisation.* | No submission or attempt. (0) | The image supplied contains the incorrect graph type or the incorrect variables are used. (2.5) | The image supplied contains the correct graphs using the correct variables, but does not adhere to all instructions. (3.5) | The image supplied contains the correct graphs using the correct variables, and adheres to all instructions. (5) |
| **Question 4: Interpreting the visualisation**  *The submission correctly interprets the Tableau visualisation.* | No submission or attempt. (0) | The student attempts to interpret the Tableau visualisation, but shows little to no insight. (2.5) | The student interprets the Tableau visualisation, but lacks some insight. (3.5) | The student interprets the Tableau visualisation and shows substantial insight. (5) |
| **Question 4: Interpreting the final model**  *The submission correctly interprets the final logistic regression model by describing the nature of the relationship between the dependent and independent variables, and provides insight into how the relative importance of each variable can be ranked.* | No submission or attempt. (0) | The student either correctly interprets the final logistic regression model or provides insight into how the relative importance of each variable can be ranked, but not both. (2.5) | The student correctly interprets the final logistic regression model and provides insight into how the relative importance of each variable can be ranked. However, the insight offered is limited or incomplete. (3.5) | The student correctly interprets the final logistic regression model and provides insight into how the relative importance of each variable can be ranked, showing substantial and comprehensive insight. (5) |
| **Question 4: Model fit**  *The submission provides an insightful argument regarding the fit of the final logistic regression model.* | No submission or attempt. (0) | The student demonstrates substantial engagement with evaluating model fit, but comes to the incorrect conclusion. (2.5) | The student demonstrates substantial engagement with evaluating model fit, and comes to the correct conclusion. However, the reasoning lacks insight. (3.5) | The student demonstrates substantial engagement with evaluating model fit, and comes to the correct conclusion, providing insightful reasoning. (5) |
| **Question 5: Decision rules**  *The submission provides a list of the six correct decision rules.* | No submission or attempt. (0) | The student provides a list of decision rules, of which three are correct. (5) | The student provides a list of decision rules, of which four or five are correct. (7.5) | The student provides a list of six correct decision rules. (10) |
| **Question 5: Evaluating the model**  *The submission insightfully argues whether the model generalises well to unseen data.* | No submission or attempt. (0) | The student addresses whether the model generalises well to unseen data, but the reasoning is partially incorrect. (5) | The student correctly addresses whether the model generalises well to unseen data, but the answer lacks insight. (7.5) | The student correctly addresses whether the model created generalises well to unseen data and provides substantial insight. (10) |
| **Conclusion**  *The submission provides a summarised list of decisions that could be made by the business based on the results of the analysis.* | No submission or attempt. (0) | The student provides a summarised list of decisions that could be made by the business, but these decisions lack substantial insight. (5) | The student provides a summarised list of decisions that could be made by the business, but the reasons for these decisions could be expanded. (7.5) | The student provides a comprehensive summary and list of the decisions that could be made by the business that shows substantial insight. (10) |

**Total:** 120 marks