Name: Yifan Zhang
Student id: s4686474

DATA7001: Introduction to Data Science, S1, 2023
Case Study Assignment

For this case study, I decided to consider Reports 3 and 6.

Report 3: Unemployment In Queensland

A. Why was this analysis carried out? What value can it bring to stakeholders? Who are the stakeholders?

A study of this issue will allow the Queensland government and the public to understand why this problem is occurring and will allow the government to put policies in place to prevent possible problems and allow the people to know how long the effects of unemployment will probably last.
For the different stakeholders, they can adjust the arrangement in time and make appropriate and positive response.
1. Individuals: This analysis can be used to understand the situation and trend of the job market and make better choices.
2. Government: Use this analysis to make more efficient policies to promote economic development and provide job opportunities.
3. Enterprises: to better understand the supply and demand in the job market and to adjust their recruitment plans.
The main stakeholder may be working or looking for work in Queensland, the Queensland Government and various businesses are also stakeholders.

B. What was the form and format of the data used? What can you say about the data size and specific format?

The main forms of data used are numeric data, image data, time series data and textual data.
For the format of the data used is TXT, Excel (charts).
For the size of the data: I think a large amount of numerical data was collected, perhaps more than ten thousand relevant data.
For the format of the data: I think it is the year, month, number of people, age of people, etc., to unify the data to be processed and to make it easy to integrate the processed data.

C. Based on what you can understand – what tools were used in the analysis? You don't have to understand all of the tools used – but look some of them up to see what is the purpose?

R was used.
R: Checked data integrity and processed missing data. Filtered out the requiring data and used graphing tools to make the processed data clearer and more concise.
Used line analysis and logistic regression analysis to analyze the data.
Line analysis: Line analysis was performed on three different sets of data, and r-squared was used to determine which set of data was more correlated.
Logistic regression analysis: logistic regression was performed by setting up two different groups to determine which group had the larger percentage.

D. How can such analysis affect decisions? What are important questions that need to be solved?

Through the data analysis, it can make the decision makers more accurate to know the future trend and development direction, so it will have more positive and useful influence in making the decision.
Through this analysis, we can conclude that the unemployment rate is not only affected by COVID-19, but also by the decisions made by the Queensland government, so what needs to be improved and addressed is how to make the government make policies that are more conducive to development and employment.

E. Imagine you were speaking to stakeholders associated with these datasets. What questions would you ask?

Do you understand what has led to your current situation and employment environment?
Do you know the future direction of employment and how to survive during the affected period when getting conclusions from this analysis?

Report 6: Example Title of Report 4

A. Why was this analysis carried out? What value can it bring to stakeholders? Who are the stakeholders?

The study of this question can help to understand the scoring of a team in a game and the accuracy of players' shooting and enable fans to better predict the trend of the game.
It allows a team's management to adjust the team structure and team tactics to improve the team's winning percentage. For the players, helping their teams to score more points will increase their salaries. And fans and sports gamblers can more accurately predict which team has a better chance of winning.
The main stakeholders are the team owners, executives, coaches and players. Fans of the team and sports gamblers are also stakeholders.

B. What was the form and format of the data used? What can you say about the data size and specific format?

Main forms of data used: numerical data, image data, time series data.
Main data format: CSV
Size of data: 20 CSV files, each with 31 rows
Format of data: The first 30 rows of each file represent the teams in the NBA, while the last row is the average of the whole NBA season.
The data is handled in this way to make the huge data structured and to clearly reflect the individual teams' shooting scores, which is more helpful to analyze the data.

C. Based on what you can understand – what tools were used in the analysis? You don't have to understand all of the tools used – but look some of them up to see what is the purpose?

Using R, 20 CSV files were processed to get the data with team, league and season scores, and the data was processed as needed to get the final results.
Linear regression (2 features), k-means clustering, and linear regression (stepwise selection) were used.
Linear regression (2 features): by creating two variables (3-point percentage and percentage of 3-point attempts) to determine the relationship with Point Differential before.
k-means clustering: The team and season 3-point percentage and the percentage of 3-point attempts are used as different clusters, and then the team or season corresponding to the most appropriate clusters is analyzed.

Linear regression (stepwise selection): roughly the same as linear regression (2 FEATURES), but this method focuses on whether teams will be successful because of the percentage of three-point attempts made, and 13 FEATURES are selected from 17 FEATURES for the final model building.

D. How can such analysis affect decisions? What are important questions that need to be solved?

I think it will affect the corresponding game tactics made by team management, and the team management plan. Training the players so that their shooting accuracy increases instead of shooting more 3-point.

The conclusion from this analysis is that there is no one right combination of 2-point and 3-ponit, and that the best improvement is in shooting accuracy rather than in the number of points shot.

E. Imagine you were speaking to stakeholders associated with these datasets. What questions would you ask?

For the team management and coaches how can they improve the training efficiency while ensuring the players' normal work schedule so as to improve the team's shooting percentage?