Name: Jaskeerat Singh

Student ID: s4761003

DATA7001: Introduction to Data Science, Semester 2, 2022

Case Study Assignment

For this case study, I decided to consider Reports 2 and 3.

Report 2

A. **Why was this analysis carried out? What value can it bring to stakeholders? Who may be stakeholders?**

The analysis was taken out to give details regarding the various factors like covid 19 pandemic, age, and education that affect the unemployment rate in Queensland. Going through the report, the key stakeholders that have been identified are the people currently working or seeking jobs in Queensland. These can include job hunters who can find this information useful for a job change and students who can use this information to make better decisions and prepare for their future careers. The other entities/stakeholders include the industries that are affected by the non-availability of workers and the government institutions that issue epidemic control in the country and hope to improve these policies for the citizens to boost employment during the covid 19 period.

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

The data used in the analysis comes in the form of structured data from CSV and Excel files. The data is made up of numerical and categorical columns. The labour force summary excel file has been divided into different sheets based on age and the time period has been set as the index for the data. The data was huge in size and therefore was trimmed down for the requirement of the analysis. As the data had entries from 1990 to 2022, the data after 2019 was selected as the employment rates needed to be compared from 2019 to 2022 covid time. The labour employment excel contains data of different regions in separate sheets based on various factors. The main “Employment PT-FT” sheet contains information on different regions and employee statistics for part-time and full-time work. The data is in the form of 76 rows x 4 columns and is in the form of string, dates and numerical values. The information about the covid case location dates and location is in a CSV file and consists of 7 columns and approx. 700,000 rows of data. All of the data is in string format.

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

EDA was performed on the data with visualization/plotting tools like tableau to check if any relationship exists between the data. it is also used to gauge and check if there has been an increase or decrease in the values of different features/attributes of the data. Linear analysis was performed between various variables to establish a relationship and a trend between the variables. here the p-value which is a statistical measurement is used. it is used in hypothesis testing and helps in determining if we can reject the null hypothesis. A null hypothesis defines whether there is no relationship between our chosen variables. The p-value tells us that if the null hypothesis is true then how often can we see a test statistic as extreme as the one calculated by a statistical test. Its value will never be 0 or 1 but the lower the value better it is. R-squared is another measure which is used in the analysis. It explains the movement of the dependent variable based on the movement of the independent variable. its value ranges from 0 to 1 and can be represented as percentages. It is the basic goodness of fit of data. A low r-squared value is bad for prediction models.

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

Such an analysis can affect the decisions of the stakeholders. Analysis of data shows that the covid 19 pandemic did not affect unemployment in Queensland. The graphs show the different fields in which men and women apply for. This can serve as recommendations for the students who are searching and making plans for a career in such fields. They can tailor their studies according to the career they want to pursue. Job seekers can start looking for jobs in other industries and government can give a boost to low-performing, unappreciated industries. This can open a range of opportunities for job seekers and students. From the analysis it is clear that in old age employment tends to go down. Government can work with old people to fill vacancies in underperforming industries.

- Has covid affected any industries which used to depend on manual workers?
- What education level has the least unemployment?
- Why is there a steady decline in the employment as the age increases?
- What factors affect job seekers and students while joining a particular career track?
- Has the pandemic made it tough to change/find jobs for the job seekers?
- What age and education level affects family relationships while pursuing a full-time employment?
- What education level or program needs to be promoted to boost employment in dwindling industrial sectors?

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

- Comparison of employment w.r.t covid 19 cases should affect after some time not immediately, so should analysis be done after taking a gap to make better predictions?
- Does the location of a job affect the decision of people to take the job?
- Among education, age and family what deters a person in taking up a job?
- Does age play a role in some individuals taking up education which can help in employment?
- Does unemployment affect an individual's decision to take up a job in another profession?
- Has the pandemic made it tough for job seekers to change jobs in their profession?
- Does a high level of education in a field make it tough to find jobs in other fields?
- Does the government have programs or policies that can help students after studies?

Report 3

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

In this project, the analysis was taken out to find out what category of crime dominates different regions of Queensland. The key stakeholders can be considered as ordinary citizens residing in or planning to move to Queensland. They can use the results as a guide for avoiding dangerous zones while buying property. Businesses operating in Queensland can use the analysis to set up shops in locations with low crime. This will also help in defining the operating time of the businesses. Law enforcement in the state can use the results to deploy manpower according to priority and level of crime. Queensland state government can use the results to make policies and use public funds better by conducting drives or operations to make the neighbourhood safe.

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

The data used in the analysis comes in the form of structured data from CSV and PDF files. The data is made up of numerical and categorical columns. The offence rates CSV file contains data contains monthly data from 1997 to 2022. The file consists of 89 columns covering a variety of offences of different severity levels. Majority of the data is in a numerical format where a combination of month and year serves as the index for the data. The offence number CSV file contains data about different regions of Queensland. The file consists of 91 columns covering a variety of offences of different severity levels. The data is a mix of numerical and categorical data. The annual report contains information about the budget and the manpower allocated to different regions in the state. The original data was cleaned of all the duplicates and monthly data was rolled into yearly data to get a compact view of the data. The data was cleaned to convert string columns into numerical values and various offences were rolled into different categories.

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

EDA was performed on the data with the use of visualization/plotting tools like tableau to check if any relationship exists in the data. It is also used to gauge and check if there has been an increase or decrease in the values of different features/attributes for different periods of time. Python has been used to develop a linear Regression model to predict the number of assault cases in the year 2022 with the same level of police manpower and budget. It also presents an r-squared score, which is a measure that shows the movement of a dependent variable based on the movement of the independent variable. Its value can be between 0 and 1 but can also be represented as a percentage. A low r-squared value is considered bad for a prediction model. In the analysis, a prediction model is plotted between different years and the total number of cases. An r-squared score of 0.6 is also produced.

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

After the analysis, various graphs have been produced to establish a relationship between various features/attributes. From the analysis, it is apparent that even after increasing the police budget the total number of cases of various crimes has increased throughout the years. Law enforcement can use this data to divert attention to other crimes whose cases have kept on increasing. The state government can use this data to increase police manpower in the regions/districts where crime has increased over the years. Citizens can use this analysis to buy properties in the districts where the crime density has not increased much from 2011 to 2021. Citizens can also use this data to start families in the regions which have a low number of serious crimes. Businesses can use the data to open shops in the regions where crime is low. This data can be used to get insurance that covers specific crimes against their property.

- Why over the years the number of offences committed remained constant if the police budget has gone up ?
- How much police manpower needs to be increased to help business owners as the number of offences against property or stolen goods has increased over the years?
- In what regions/districts does the state government needs to increase their activities to reduce crime where the crime density has not decreased in a decade?
- How are the state government and law enforcement planning to tackle districts with a high rate of crime, as an increase in police budget and manpower has failed to lower the crime spree?

- What strict measures needs to be taken to reduce crime related to assault and harassment?

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

   - What major crimes affect the decision of an individuals to move to a region or district?
   - How much of a increase in police manpower is needed to bring down the number of assault cases?
   - Even after increasing the manpower and police budget, there are no changes in the dangerous zones even after a decade. what can the state government introduce to public to bring down the offence rate?
   - What affects more in a business decision to open shop in a particular region crimes vs property or crimes vs individuals?
   - There is a sharp decrease in the value of stealing and stolen goods offences. An introduction of covid 19 lockdown data most probably can show a good linear relation for these type of offences.