**CP5046 Assessment Task 1 – Project Documentation**

This assessment task has been prepared by Dr. Dmitry Konovalov for James Cook University.

© Copyright 2020

This publication is copyright. Apart from any fair dealing for the purpose of private

study, research, criticism, or review as permitted under the Copyright Act, no part may

be reproduced by any process or placed in computer memory without written permission.

**ASSESSMENT TASKS**

|  |  |
| --- | --- |
| Aligned subject learning outcomes | All |
| Group or individual | Group |
| Weighting | 20% |
| Due date | Week 7, before the workshop |

**DESCRIPTION**

This task is the **iteration-1** in terms of the Agile Software development. It defines the initial project specifications including goals, deliverables, and planning for **iteration-2 (alpha release)**.

**CRITERIA SHEET**

[\_\_\_\_\_/20 marks] Assignment is done in a group with 2-4 students. Every team member gets the same mark for this assessment item Write group members here:

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Project role** | **Project contribution** |
| Vikash Singh | Project Manager | Responsible for setting up GitHub repository, scheduling recurring meetings and delegating tasks. |
| Willis Lin | Document officer | Responsible for data gathering, version control of the documents and testing of the solution. |
| Kevin James | Data Analyst | Responsible for dataset pre-processing using Weka and unit testing. |
| Harish Shetty | Development & Support Engineer | Responsible for writing Python code in PyCharm, finding patterns for road crashes and unit testing. |

[\_\_\_\_/Prerequisite for marking] Assignment is completed using electronic copy of this document and submitted to LearnJCU electronically. One submission per team.

[\_\_\_\_\_/20 marks] Project description for non-ICT-technical stakeholders and general audience. Write here: minimum TWO pages, maximum TEN pages.

• [\_\_\_/10] Justification for the project: Why a new ICT solution is required. Include some market and ICT technology research, plus available ICT solutions.

|  |
| --- |
| The aim of our project is to discover patterns on road crashes and number of persons killed yearly.  We will be analysing the dataset provided by the Bureau of infrastructure, transport and regional economics (BITRE, URL: <https://data.gov.au/dataset/ds-dga-5b530fb8-526e-4fbf-b0f6-aa24e84e4277/details?q=road%20crash>) - a government body of Australia, using data mining techniques to generate common causes for road crashes in several categories such as speed limits, gender, age and holiday periods, etc. The complete report will be provided to the client for the improvement of Australia Road Safety and Regulations.  Reducing traffic accidents is an important public safety challenge around the world. Accident pattern is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer.  The current practice is to store road crash data into a csv file and use excel to analyse the data. It is efficient when the dataset is small, but it becomes less effective and problematic when the dataset is large and contains missing fields. Some ICT related research has been conducted using clustering algorithms and association rules to discover patterns that cause road crashes and to identify risk factors that affect road crashes. Available ICT solutions include Hazardous Traffic Event Detection, warning of speed limits via GPS, notification of hazardous road condition, just to name a few. By using data mining techniques to analyse the dataset, we can find meaningful patterns to prevent road crashes and to enforce policies and regulations for Australian Road Safety. |

• [\_\_\_/5] Project goals: Describe exactly what and how your proposed ICT solution will be delivered to the client.

|  |  |
| --- | --- |
| Goal Number | Goal Description |
| 1 | Exploratory Data Analysis (EDA):  Pre-processing dataset, get rid of noise data using Python on Kaggle notebook. |
| 2 | Exploratory Data Analysis (EDA):  Extracting important variables to uncover hidden structure in order to maximize insights of the dataset. |
| 3 | Analysing dataset to find patterns of causes and effects. |
| 4 | Developing reports of findings, including graphs and charts. |
| 5 | Summarizing the project and uploading it to Kaggle notebook. |
| 6 | Providing project findings to the client for future improvement of Australian Road Safety. |

• [\_\_\_/5] Justify between two and four major milestones with timelines. Is your proposed schedule too ambitious (over-optimistic) or too conservative?

|  |
| --- |
| 1. Load data to Kaggle and clean up noise data (26 days). 2. Design initial feature on Kaggle (11 days). 3. Develop several charts and graphs to identify focuses on interesting patterns   (11 days).   1. Finalise reports and design a user manual (16 days).   The first milestone is on dataset pre-processing using Kaggle. We are optimistic that we can clean up the dataset and extract important information within 26 days. The second milestone will be designing the initial features on Kaggle according to the information extracted from the dataset which we believe that it would take around 11 days. The third milestone is to develop charts and graphs so that we know which one we should be focusing on for more insights from the dataset. It is estimated that we would spend approximately 11 days on this. Lastly, in milestone 4, we will be finalising our report and designing a user manual for our client to facilitate the understanding and utilisation. It is anticipated that we will be spending a considerable time on this, which will be around 16 days. We believe if we follow our schedule each week in doing the work collaboratively, we will achieve our goal on time. |

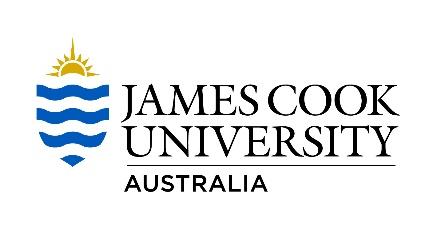
[\_\_\_\_\_/20 marks] Project scope of the full final release (Project audit at the end of CP5047 subject). Clearly specify the scope of alpha release (Project audit at the end of CP5046 subject), and. Write here: Minimum TWO pages, maximum TEN pages. Available days: PASS-level (minimum) effort is 1-day and maximum are 3-days per teaching week (assume 12 weeks per study period), per team member. For example: 12 x 2 = 24 days is the minimum for a team with 2 students, 12x2x3=72 days is the maximum. The available days must match the SUM of your user story estimates below. User story 1: title, short description, effort estimate (in days, maximum of 5 days). User story 2: …

Available days: 12 \* 4 \* 1 = 48 days (Minimum) 12 \* 4 \* 3 = 144 days (Maximum)

|  |  |  |
| --- | --- | --- |
| **User Story Title** | **User Story Description** | **Effort Estimate** |
| Data Cleaning | Data can have a lot of missing values or irrelevant parts. To handle these parts, data cleaning needs to be done for missing and noisy data. | 5 days |
| Death by age group | To provide accuracy in the death rate by age we have populated analysis between the age and crash year attribute | 2 days |
| Deaths by day of the week | We have analysed the death rate based on the day in a weekday/weekend to find the threshold death’s by day | 2 days |
| Pedestrian fatalities | To reduce the minimal death’s, we have analysed the count of deaths by pedestrians (we can compare with region) | 2 day |
| Vehicles involved | We have compared the involvement of various types of vehicles with the road deaths to find the majority rate. | 5 days |
| Number of accidents state-wise | Representing accident rate in each state | 3 days |
| Number of accidents by gender | To find the fatality rate pattern we have analysed accidents by gender | 5 days |
| Number of deaths state-wise | Representing death rate in each state | 5 days |
| Accidents in different speed zone | Accident rate on each speed limit zones | 5 days |
| Accident rate during the day | We analyse the dataset to understand the percentage of accidents happened during the day in each state. | 4 days |
| Accident rate during the night | We analyse the dataset to understand the percentage of accidents happened during the night in each state. | 5 days |
| Fatality rate by road user | We analyse the dataset to uncover the percentage of death by road users, including passenger, driver, cyclist, motorbike rider. | 5 days |
| Crash type | We analyse the dataset to find the percentage of each crash type for pedestrians, single, multiple annually in each state. | 5 days |
| Creating the report | Creating a report through analysing the dataset, obtaining meaningful patterns, illustrating graphs and charts to assist the client’s understanding. | 5 days |
| Creating user manual | We will create a user manual for the client so that the client knows how to interpret the program and result. | 5 days |
| Final Submission | Upon successful validation we submit kernel documentation for open source. | 1 days |

Total number of days spent on this project is 64 days.

[\_\_\_\_\_/20 marks] Project sponsor/client/customer signed/agreed to the scopes of the alpha-release and the final-release. Cut/Paste clients’ email here, or show signed the preceding “Project scope” to your marking lecturer.



**Client Consultant Engagement**

For ICT -1

Between: **Dimitry Konovalov*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

(including its successor ‘the client’)

and: **Kevin Vathalloor James, Sai Harish Dachepalli, Vikash Singh, Wei-Chih Lin (Willis)*\_\_\_\_\_***

(including its successor ‘the consultant’)

|  |  |
| --- | --- |
| ***Project:***  ICT - 1 | ***Location:***  JCU Townsville |
| ***Client’s Representative****:*  *Name:* Mr. Dimitry Konovalov  *Contact Details:* [Dmitry.konovalov@jcu.edu.au](mailto:Dmitry.konovalov@jcu.edu.au) | ***Consultant Representative for the group****:*  *Name:* Mr. Vikash Singh  *Contact Details:* [Vikash.singh@my.jcu.edu.au](mailto:Vikash.singh@my.jcu.edu.au) |
| ***Scope and nature of the Service***:  Analysing the dataset provided by the Bureau of infrastructure, transport and regional economics (BITRE) - a government body of Australia, using data mining techniques to generate common causes for road crashes in several categories such as speed limits, gender, age and holiday periods, etc. The complete report will be provided to the client for improvement of Australia Road Safety and Regulations. | |
| ***Information and Service to be provided by Client*:**  Specifying the requirements and providing details for the source and data required for analysis and publication online. | |
| The Client engages the consultant to provide the service described above and the consultant agrees to perform the services (including as may be set out in any relevant statement of work provided to the consultant by the client). Both parties agree to be bound by the provision of this client consultant engagement. Once signed, this agreement will replace all or any oral agreement previously reached between the parties. | |
| ***Client Authorized Signature:***  ***Print Name:*** Dimitry Konovalov  ***Date:*** | ***Consultant Authorized Signature:***  ***Print Name:*** Vikash Singh  ***Date:*** |

[\_\_\_\_\_/20 marks] Project development and release ICT infrastructure. This must include development environment, programming languages, source code repositories (Configuration Management), project collaboration tools, and development tools. Write here: minimum TWO pages, maximum TEN pages.

• [\_\_\_/5 marks] Configuration Management/version control, e.g. git, github, heroku, bitbucket;

|  |
| --- |
| As part of the ICT project 1, the version control that we use in this data scenario is Kaggle. Kaggle is an online community that contains enormous datasets that can be used for data manipulation and data mining. The Kaggle can also be used to publish a dataset, explore and build models that can be helpful for accessing it to a web-based environment. Another important part of Kaggle is the Kaggle notebook which is a programming environment for data processing and report generating. It is a free platform known as Kaggle kernel that is a Jupyter notebooks in a web browser. |

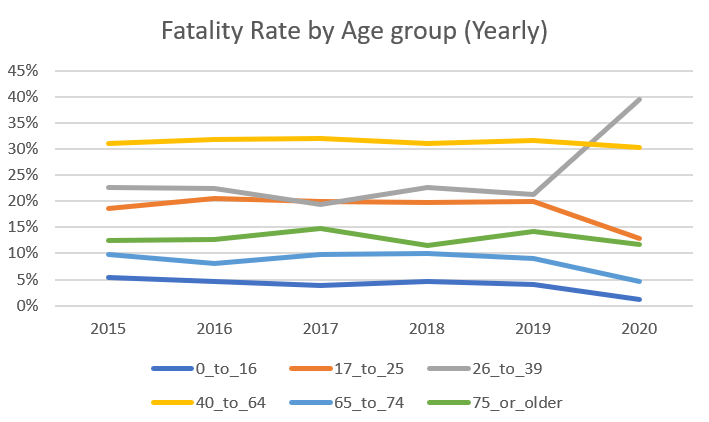
• [\_\_\_/5 marks] Project tools. Programming languages/IDEs, e.g. php/WebStorm, java/IntelliJ; Building tools/procedures, e.g. git-push to heroku; e.g. how to set-up your development environment for a new team member

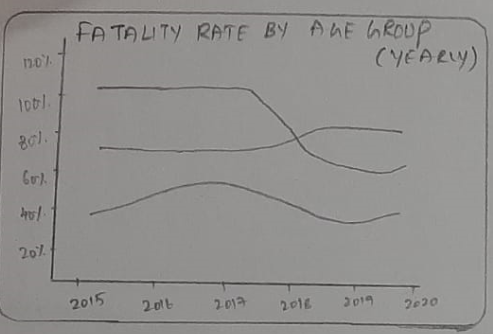
|  |
| --- |
| Developing a report from a bunch of excel files manually is a tedious process. To make reporting easier and simple, one of the best solutions is to get it done in a program that generates reports according to the parameter provided. The program that we use in this case is python 3.8. It is the newest major release of the Python programming language, and it contains many new features and optimizations. Python is one of the most popular programming languages that is used for machine learning and data exploration. |

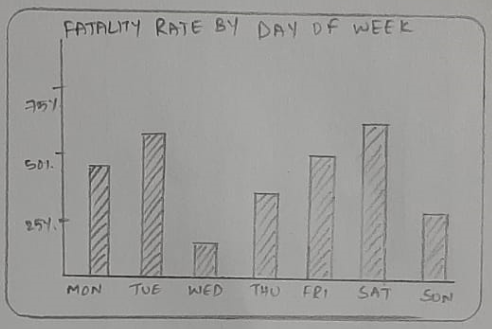
• [\_\_\_/5 marks] Testing tools, data and procedures (what and how you are planning to test), e.g. junit; Client testing and access to release, e.g. domain name, domain hosting; Client training document, and procedures; Release testing tools, data and procedures (what and how you are planning to test);

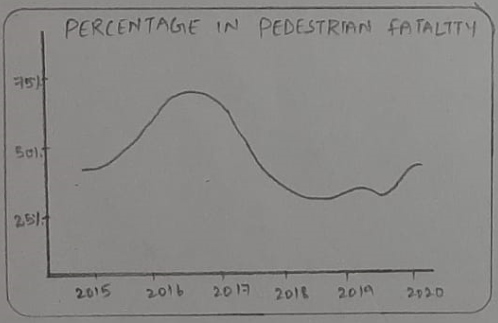
|  |
| --- |
| As part of data pre-processing, testing is one of the main processes that cannot be neglected. Testing allows feedback and errors from the user point of view that can be helpful for effective development. The functionality of the project depends on the authenticity and accuracy of data. Thus, the data processed will be the first form of testing from the developer point of view  Secondly, the processed data needs to be presented visually to the clients such that the data analysing and report generation can be done. To comply with this, the developers can create a training document on how to process the report in Kaggle. The client training document contains the process that needs to be done in Kaggle to view the content, how an API works such that live data can be accessed and general ideas about Kaggle Notebook are explained. |

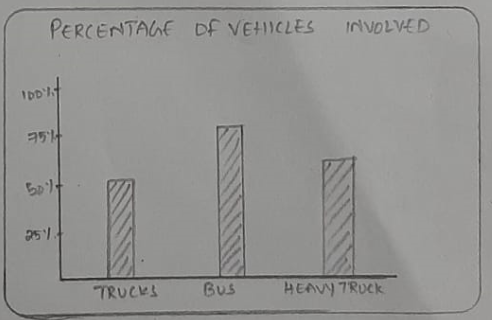
• [\_\_\_/5 marks] Prototypes are demonstrated to justify the proposed ICT solutions;

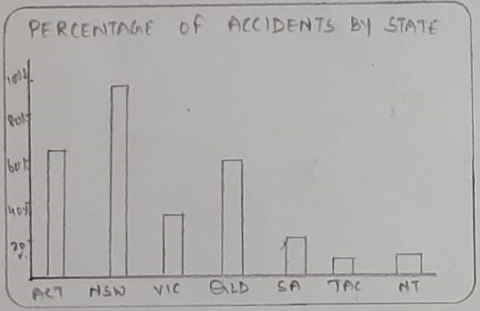


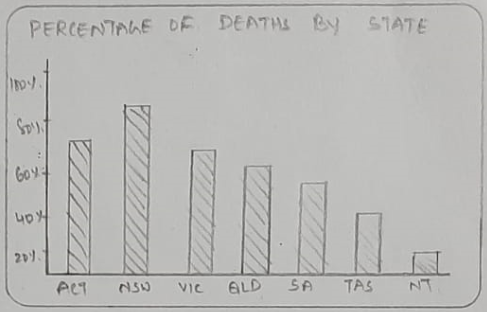


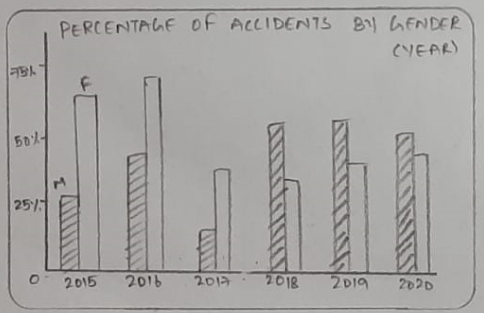


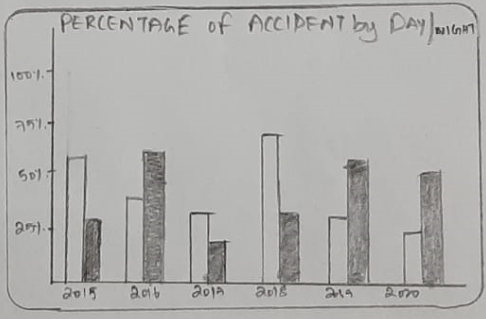


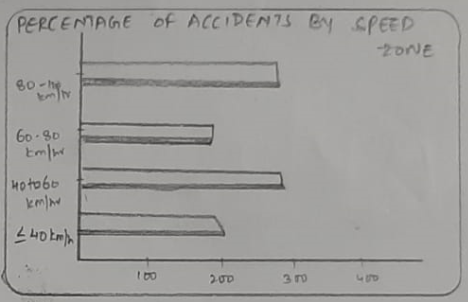


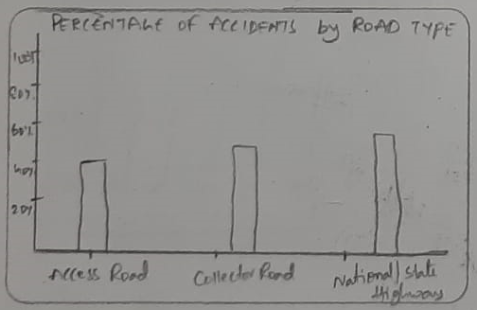












**APA Reference**

Bureau of Infrastructure, Transport and Regional Economics. (2020). Australian Road Deaths Database (ARDD). Retrieved from https://data.gov.au/dataset/ds-dga-5b530fb8-526e-4fbf-b0f6-aa24e84e4277/details?q=road%20crash

Lukuman, W., Haobin, J. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PIoS one*, 14(4), e0214966. doi.org/10.1371/journal.pone.0214966

Shuai, Y., Yuanhua, J., Dongye, S. (2019). Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States. *Sustainability*, 11(7), 1925. doi:10.3390/su11071925

Zhonggui, Z., Yi, M., Gangbing, S. (2019). Identify Road Clusters with High-Frequency Crashes Using Spatial Data Mining Approach. *Applied Sciences*, 9(24), 5258. doi.org/10.3390/app9245282