7CCSMDPJ

Individual Project Report 2017/18

Student Name: Luke William Ronald Whittaker

Student Number: 1763234

Project Title: London Traffic Matters

Degree Programme:

Date of Submission:

Supervisor Name(s): Dr Rita Borgo

Word Count:

This dissertation is submitted for the degree of MSc in Data Science.

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

I **agree** to the release of my project

I **do not** agree to the release of my project

Signature: Date:

Abstract

* The Abstract is a short, executive summary of your project. This should include a brief description of the project objectives and research question(s) addressed, followed by a brief description of the main contribution(s) of the project, including a summary of the results achieved and the primary conclusions drawn from the work. This should appear on a single page by itself and should be the second page of the report.

Acknowledgments

I wish to thank my project supervisor Dr Rita Borgo whose help and guidance was beneficial to the completion of this project.

Table of Contents

This is a list that includes and identifies the main sections, sub-sections and appendices (if any) of the report.  Page numbers must be included.

Nomenclature

1. Introduction 6

2. Background 7

2.1 What is Data Visualization 7

2.2 Visualizing Data on Maps 8

2.2.1 Map Projections ?

2.3 Different Software Technologies ?

2.3.1 Python ?

2.3.2 JavaScript ?

2.3.3 WebStorm ?

2.3.4 XAMPP Server ?

2.3.5 Leaflet JavaScript ?

2.4 Glyph Visualizations ?

3. Related Work 8

4. Approach 10

5. Results 11

6. Conclusion 12

7. References 13

8. Appendices 14

8.1. Appendix: Hints for Success 14

8.2. Appendix: Submission Details 16

8.3. Appendix: Plagiarism Warning 17

Glossary

This is a glossary defining all abbreviations and symbols used in the report. The contents should be listed in alphabetical order. Note that placing entries here is NOT a substitute for properly and completely defining terms in the body of the report, accompanying the first time a term is used.

*SVM* Support Vector Machine 7

DfT Department for Transport 6

GB Gigabyte 8 these will need to be changed

GIS Geographic Information System 8

TB Terabyte 8

List of Figures and Tables

These are lists of all the figures and tables found in the report.

# Introduction

The Introduction is the first content section of your report. You should describe the general area (e.g., application domain) in which your project research is conducted, the motivation for conducting the research and the overall aims of the research. Be sure to outline your research questions and give a brief summary of the conclusions drawn, though the conclusions will be detailed later in the report. With the Introduction, you want to interest your reader and tell them why they should care about your research and why they should read the rest of the report. The report will be read (marked) by examiners with a technical Computer Science background, but not necessarily any knowledge of your domain, so make sure that you provide enough information for a naive reader.

# Background

Visualizing data is a vital process to undertake when exploring and aiming to bring insight to any format of data. With this in mind there are some aspects that should be taken in to consideration when exploring the data. As mentioned previously the data to be used for this project is traffic data provided by DfT. Thus, we must ensure that the data visualizations used to explore and analyse this data are appropriate for the tasks performed. The aspects that should be consider are: choosing the appropriate graph or visual encoding to bring meaning; this can take the form of the most appropriate graph to show the relationship between two variables, one having a numeric data type and the other taking a nominal data type; in this instance a Bar chart would be appropriate to show this relationship. Moreover, when it comes to deciding what map graphic to use when handling geo location data; we must look at the type of projection of the geology (this is be explored further in Map Projections) and the aesthetics of the map, this could be similar to Google’s satellite view, or the default view in Google Earth and Google Maps respectively.

In addition, the visual encoding used is of importance as this allows the reader to understand what the data means. Thus, if we were to have visual encodings that do not appropriately display the results then incorrect or misleading conclusions may be formed. Thus, this section of the report will look to explore in detail the theory behind data visualization, how we can visualise data through maps, technologies available to aid the creation of such visualizations, and advanced ways to enhance visual encodings that provide a greater level of insight to the data.

## What is Data Visualization

Within data analytics there are many sub-domains that exist to provide standardized ways in which to extract, transform, process, and display any format of data available. Data visualization is one sub-domain of data analytics; however, its main focus is how to best display data. With this in mind, we can deem it a vital domain of analytics due to its role for presenting data in such a way to bring instant understanding upon first sight. Moreover, with this visual representation of data, we are able to draw critical insight in to specific patterns unique to the data. Additionally, we are able to convey complex concepts in ways that enable readers of all levels to be able to gain insight to the data.

The concept of using visualizations to understand data has been used by humans to good effect for many centuries, whether this was the Romans using maps to navigate throughout the Roman Empire, to John Venn who inventing the Venn diagram that visualises symbolic logic. However, modern day visualizations still draw from past techniques, such as bar charts for visualising data with features having numeric and nominal data types. Further, bar charts can be enhanced through the use of colour to distinguish between the different nominal features. Additionally, line graphs are still well suited to visualizing temporal data – data recorded over a certain period of time – and many more visualizations. Moreover, with the technology age it brought with it the ability to take these visualizations to greater heights; this can be seen in the amount of data that can be processed, with modern computing we are able to process high volume of data, data which can range from 1-10s of GBs to many TBs of data.

Having this ability to visualize large amounts of data, for example stock market trend data visualized as a line graph, we are able to make quicker judgments on whether to sell, buy, or hold on to stocks. In contrast, if we were to view this data in a Microsoft Excel spread sheet, we might not be able to intuitively see what the trend might be at first glance. Moreover, such data representing the stock market would be obtained on a constant basis, which would result in the data held within this spreadsheet being time intensive task to scroll through to try and find patterns. In addition, handling data regarding Geographic Information Systems, more commonly known as GIS data, has been greatly enhanced with modern computing. This level of computing allows for processing of 10,000+ data regarding geographical locations, which leads on to the discussion regarding visualizing data on maps.

## Visualizing Data on Maps

### Map Projections

## Different Software Technologies

During the initial stages of this project, decisions needed to be made regarding the process and the technologies to be used throughout the creation of this project. Within this section discussions will be had exploring the pros and cons to each of the technologies used and explored throughout the completion of the project.

### Python

Before we can discuss what aspects of the project would be aided through the addition of Python, we must first look to understand the motivation and characteristics of the Python programming language. Python is one of many human-readable programming languages, more specifically called higher-level languages. In addition, Python programmes are processed through the Python interpreter; this way of program execution varies from many over languages, such as a language called C#, pronounced C-Sharp. C# is what known as a compiled because it is compiled in to machine language code before it is executed; we will speak more about this later.

Python was created by Dutch Computer Scientist Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands [?]. Python takes a lot of its functionality from many languages, more specifically C and C++; we can see this through the Cython interpreter, which allows for C extensions to be built with the ease of writing Python programmes but with the speed of C.

Now that we know who created Python we can explore the fundamental philosophy behind its creation and see how it differs from other languages. Firstly, Python aims to ensure that function, class, or loop can be easily read by any reader; it is this level of readability that makes Python one of the best programming languages for beginners to learn. With this in mind we can take a look at the five philosophies of Python, which include:

* Beautiful code is better than ugly code
* Being explicit if much better than being implicit
* Simple code is better than complex code
* Complex code is better than complicated code
* Readability is important

Python looks to apply all of these philosophies by using English instead of symbols for logical operators [?]. For example, in C like languages we would use “&&” for the logical operand AND; whereas in Python we use the English word “and”.

When comparing Python to other high-level programming languages we can see there are some distinct differences. These differences come in the form Python being a strongly typed and dynamically typed language. Python differs from a language such as C, which is a weakly typed language [?]. For example, in C we can have the following legal code:

byte foo = 50;

int f = foo;

The reason the above is valid, is because the C compiler will make the type conversion from a byte to an integer for us. However, if we were to perform the following in Python it would result in an error message appearing:

foo = 50

f = “40”

print foo + f

This would not be accepted by the Python interpreter because we are looking to add a number and a string together, in order to successfully execute this, we would need to convert the string in to a number.

As mentioned above, Python is a dynamically typed language, this means that the Python interpreter does not know what data type each variable has before the whole program is executed. This allows for some level of freedom with variables; however, it does mean that all variables must be instantiated when created. In contrast, C# is a compiled language, this means that C# programmes must be translated in to machine code, this is a low-level programming language, before it is executed. One benefit compiled languages have over dynamic languages is that the compiler knows the data type of each variable before execution, thus allows for optimizations to be made to improve performance [?].

Now that we have a good understanding of what Python is as a programming language and discussed its characteristics. We can now mention how Python will be used throughout this project. Python will be used extensively throughout the data pre-processing and data curation process. The reason being that Python has libraries that aid these tasks, these include: csv, and json which aid in handling such data, and pandas which will aid in manipulating the curated data. Python was also chosen for this project due to its ease of use which will enhance the progression of the project, along with it being cross platform which meant that curation of the data would not be limited to a certain hardware.

### JavaScript

### WebStorm

### XAMPP Server

### Leaflet JavaScript

Then speak about Glyphs if I get on to this section

# Related Work

The Related Work section of your report should provide a review of recent literature in the area of your research. This is distinguished from the Background section because it is typically newer and more experimental. If there are standard terms or techniques mentioned in the literature, then you can define what these are in the Background and use the Related Work section to explain how researchers have used the standard techniques as benchmarks or fundamental methodologies for their research. For example, if you review an article that describes using k-means clustering for finding appropriate groups of patients with similar sets of symptoms, then you could describe what k-means clustering is in your Background section and describe how researchers used that technique on patient data in your Related Work section. When you review literature, be sure to explain how the articles you cite are relevant to your project. Be critical---outline pros and cons of the work you are reviewing. Be clear to explain how the work you review is different from your own work. Note that you may find it easier to compare and contrast others' techniques with yours later in the report, after you have explained your own work. That is fine---just be sure to forward reference in the Related Work where you will compare to your own work (and backward reference in the later sections back to the Related Work). This can include information that you had in your Project Proposal report that was due in April, but should typically be substantially expanded from what you had in your proposal.

Refer to links and resources on the KEATS page to help with your literature review.

Be sure to provide complete references when someone’s work is mentioned.

Examples of articles we might cite are [3] and [4].

# Approach

The Approach section of your report should describe what you did. You should discuss your research questions in detail here, explaining for each question how you addressed each question (i.e., what techniques you used) and how you evaluated the success (or failure) of your investigation. This should include a description of the data set(s) that you used for your research (e.g., what you included in your Data Acquisition report that was due in March).

This section is where you explain what you have done. Reiterate the problem you are trying to solve. Defend your reasons for choosing the techniques that you selected. Discuss the pros and cons of various existing algorithms and approaches, including narrative of things that you tried that did not work (with explanation about why they did not work). Detail the differences between your method and existing approaches in the literature (e.g., techniques you mentioned in the Related Work section). If you perform experiments, then describe the design of those experiments in this section.

# Results

The Results section of your report basically contains the answers to your research questions. This section should present the results of your evaluation, provided quantitatively, qualitatively and/or visually, as appropriate, followed by an analysis of the results. If you have performed experiments and/or analysis, then these should be presented here. Use figures and drawings to explain the significance of your results.

Discuss with your project supervisor(s) and/or domain advisor(s) how best to present your results. The main point is to make sure that it is clear to the reader what the answers to your research questions are and how you arrived at these answers.

# Conclusion

The Conclusion is the last section of your report (other than Appendixes). In this section, you can revisit the research questions and summarise your answers. Clearly explain how your investigation and your answers are a contribution---why your work is worthy of a passing mark. Also in the Conclusion section, it is good to have subsections that highlight (a) Future Work, in case you were going to keep working on the same line of research or you wanted to recommend follow-up investigation for another student to pursue next year; and (b) Lessons Learned, where you can explain how you might do things differently if you started over, because you've learned valuable things along the way (these could be technical, but they could also be personal, such as organising your time better or listening to the project coordinator who told you to BACK UP your work frequently).

# References

This is your bibliography. It is a list of source materials that you have mentioned in your report, including articles, books, web sites, software tools and libraries, data sets, videos and other items you have drawn upon to conduct your research. Any material mentioned in the report that is not your own work must be given a reference.  All references must contain the following details: *authors, title of the work, sources* (journal title, conference title, publisher (for books)), *date of publication,*and *page numbers (except for online sources)*. General guidelines for reference formatting can be found in the **Resources** section of this KEATS page.

Sample references:

1. M. Johnstone, Gaussian estimation: Sequence and multiresolution models. 2011.
2. F. Inc., “Phage lambda: description & restriction map,” November 2008.
3. J. Doe, The Title. PhD thesis, University of Mars, 2011.
4. I. Johnstone and B. Silverman, “Ebayesthresh: R programs for empirical Bayes thresholding,” Journal of Statistical Software, vol. 12, no. 8, pp. 1–38, 2005.

# Appendices

Supplementary materials may be included, such as additional tables and figures that would detract from the narrative if you included them in the main sections, above. Each appendix must be labelled (for example, Appendix A, Appendix B).  All Appendices must be referred to somewhere in the text.

## Appendix: Hints for Success

* Start by creating an outline of the report, which gives you an overall structure of the report.
* Think of the text of your report as a sandwich: the "bread" is the introduction and the conclusion. Start writing the "meat" first--the inside of your sandwich. Write about your results. Then write about the methodology you used to achieve these results, the software you wrote, the libraries you integrated and the data set(s) that you explored. *Write your introduction and conclusion last!*
* Show understanding of the topic and demonstrate the contribution of the work. At least 70% of the content of the report must be your own contributions and achievements.
* Always use your own words.
* The main report and any appendices must constitute one PDF document.
* Pages must be numbered consecutively.
* Captions must be provided for all figures and tables.
* For graphs, all axes and units must be labeled (in a font large enough to be read--a good guideline is that no label in a figure should be smaller than the font in the body of the report, even when the figure is included in the report; sometimes you have to generate the PDF in order to make sure this is the case).
* Equations (or important equations), figures and tables must be numbered.
* All figures and tables must be referred to in the text.
* Units of all variables must be provided.
* Numerical values (floating-point numbers) should be displayed with appropriate precision (e.g., 2 decimal places for currency, more or less as appropriate).
* Contractions ~~shouldn't~~ should not be used.
* Check punctuation of sentences. In particular, those sentences with equations. For example, if an equation is at the end of a sentence, a full stop should be used. If sentences are comprised of multiple clauses, use commas (,) and semi-colons (;) as appropriate, in order to help the reader understand what you are trying to say.
* All variables must be defined.
* Font face of variables throughout the report (in the text, equation, figures and table) must be consistent.
* Use proper headings for chapters, sections, subsections.
* Chapters, sections, subsections should be numbered, and the same numbering system should be used throughout the report.
* It is suggested that vector and matrix variables should be in **bold** and scalar variables should be in *italics*.
* Terms and abbreviations should be written in *italics* and defined the first time they are used.
* References must be used for text quoted in the report that is not yours, as well as software and other materials (e.g., images) that you did not generate yourself from scratch.
* A standard reference format must be adopted and be consistently applied throughout the report.  General guidelines for reference format can be found here.
* Always back up your files!!

## Appendix: Submission Details

Report:

* TheFinal Report (Dissertation) must be submitted electronically via KEATS by the 4pm deadline on the due date.
* You should make sure in advance that you can upload your report so that there are no last-minute glitches. You can upload multiple times. The final version uploaded will be the one marked (all uploads over-write any previous uploads).
* Submit the report as a PDF file. There are various ways to convert .doc/.docx files into PDF. For LaTeX users, pdflatex automatically produces PDF files of good quality.
* Do not send your final report to your supervisor directly.

Source Code:

* All work on source code must stop once the code is submitted.
* Keep a working version of your source code that you can demonstrate during the Oral Presentation.
* Your examiners may ask to see the last-modified dates of your program files, and may ask you to show that the program files used in the project examination are identical to the program files submitted with your project.
* Any attempt to demonstrate code that was not included in your submitted source listings is an attempt to cheat.Any such attempt will be reported to the KCL Misconduct Committee.

## Appendix: Plagiarism Warning

IMPORTANT NOTICE:

Given the importance of the project in the degree programme, the penalties for plagiarising project work are especially severe (and include the possibility of permanent exclusion from the College with no possibility of receiving a KCL degree). Our Department has considerable expertise in detecting plagiarised work. As a student at King's, you will have read the College's statement on plagiarism and you will have already signed a declaration to state that you understand the term and agree to abide by the statement on plagiarism. If you require further explanation of the College's policy on plagiarism, then please ask your supervisor for guidance.

**Plagiarism** is when you use someone else's work without acknowledgement, which may include concepts, design, ideas, a piece of program code, a section of text, diagrams, figures, approaches, methods, results, techniques, etc.  All materials, works or contributions that are not your own must be acknowledged, using correct citation procedures (see the **Skills Training** section of the KEATS page for more information).