|  |  |  |
| --- | --- | --- |
| Interactive exploratory data analysis on the web | 15/16 | |
| Yongyan Zheng (Carina) | |  |

Contents

[Abstract 2](#_Toc441767558)

[Introduction 2](#_Toc441767559)

[Motivations 2](#_Toc441767560)

[Working Environment Setup 3](#_Toc441767561)

[Methodology 4](#_Toc441767562)

[Primary Data Exploratory 4](#_Toc441767563)

[Functions Selecting Progress 6](#_Toc441767564)

[Background foundation 6](#_Toc441767565)

[User interface 10](#_Toc441767566)

[Application Discussion 12](#_Toc441767567)

[Results 17](#_Toc441767568)

[Conclusion 18](#_Toc441767569)

[Future Development 19](#_Toc441767570)

[Appendices 20](#_Toc441767571)

[Data Classifications 20](#_Toc441767572)

[Glossary 21](#_Toc441767573)

[Key functions: 21](#_Toc441767574)

[Full Code 22](#_Toc441767575)

[References 23](#_Toc441767576)

[Acknowledgement 23](#_Toc441767577)

# Abstract

Technology has provided more forms to perform various statistical tasks such as data analysis, data exploratory and data presentation. Because of the advantages of the convenience and common usage, the idea of Interactive Media comes out. Statistics New Zealand is trying to find the key components of a good Statistical Interactive Media via this research, which is performed by building a web application of R.

# Introduction

## Motivations

Visualization is the communication of information using graphical representations (Ward, Grinstein and Keim, 2010). Because human beings are visual beings, visualization provides a more efficient and completed method to obtain and understand information. In old times, the computer graphics were widely used. However, as the technology is getting more advanced day by day, people want to obtain information in a more interactive manner. Interactive Media meets the requirements by providing the opportunities to the users to express what they want to display, how they want the result to be displayed. At the end, users shall have their self-developed but accurate understandings towards the topics they are looking at.

In this research, we are building an interactive map of New Zealand along with the 2013 census data. Such map would show the means people travel to work and their weights (in percentages) in different territory authorities within New Zealand. The most important purpose is to provide the possibility that users can visualize the differences in regards of choosing travel means between territory authorities, and at the same time, the map shall enable users select and update the parameters in a more customized manner. Potentially, users shall develop their own understandings and go further in their fields.

Data used in this research is from Statistics New Zealand, the major producer of official Statistical of New Zealand, which collects and analysis raw data, publishes useful information in general English to the public. It is efficient in helping individuals, community groups, businesses, and government agencies make better decisions.

The key definitions involved in this research are:

1. **Territory Authority**: It is either a city council or district council. There are 67 territorial authorities consisting of 12 city councils, 53 districts, Auckland Council, and Chatham Islands Council (Please refer to section 1 of the Appendices for the full list of the Territory Authorities involved in this research).

**Agenda:**

* + Excluding territorial authorities: Chatham Islands, Area Outside Territorial Authority.
  + The version used is 2011 v1.0.

1. **Population**:
   * On New Zealand level, people who are usually live in, and were present in, New Zealand on census night.
   * On Area level, people who usually live in that area and were present in New Zealand on census night.

For example, a person who usually lives in Christchurch city and is visiting Wellington city on census night will be included in the census usually resident population count of Christchurch city. (Statistics New Zealand, 2015)

1. **Main means of travel to work**: the method a person aged 15 years and over used to travel the longest distance to their place of employment on census day (for example, by bicycle, bus, walking, or jogging).

Responses indicate the main means of travel people chose on the census day and not necessarily indicate the responders’ usual mode of travel to work.

**Agenda:**

* + Excluding means: 77 Response unidentifiable, 99 Not stated.

1. **Rounding**:The source data used in this research has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.
2. **Extra Note**: some data is confidential; they are represented by “(c)”.

There are many different types of people who will use this map for specific purposes. The typical scenarios are:

1. To-graduate or newly-graduated high school students. These students are considering attending universities in other cities. Using the interactive map, students can find out if the cities match their expectations in regards to the main means of travel to work.
2. Biologists. They are designing experiments for any relevant research topics. Using the map, they will be able to make comparisons base on their understandings and findings, hence increase the accuracies and lower the risks of the experiments.
3. Politicians. Politicians have certain policies/plans that they want to submit to the authorities and hope the public will support them. Teams of the politicians then would use this map as their references/evidences to demonstrate the advantages of the new polices or plans.

## Working Environment Setup

The components involved in this research are:

1. The Operating System is using Windows.
2. RStudio, version 0.99.489.

# Methodology

This section in separated into three sub-sections. The first two sections explain the primary setup of the methodology, followed by the process of applying the target data and the selected functions. At the end, it generates the outputs and illustrates the meanings behind the results at the first look.

## Primary Data Exploratory

The source data is retrieved from Statistics New Zealand website (<http://stats.govt.nz/>). In this research, we are using 2013 census data because it is the latest census data available as in November 2015. We can see various online tables grouping the census data by different parameters (follow Steps 1 to 5 below). Given that we are not modeling any relationship against any parameter in this research, we choose the plain table, which does not involve any parameter, by following steps below.

1. Go to **Statistics New Zealand homepage** (<http://stats.govt.nz/>).
2. Go to **Quick links** section.
3. Click on **NZ.Stat** (<http://nzdotstat.stats.govt.nz/wbos/Index.aspx>).
4. Click on **2013 Census**.
5. Click on **Transport and communications**.
6. Click on **Main means of travel to work, 2001, 2006, and 2013**.

After the steps above, it shall display the online table with original setup.

To customize the selections, do the following:

1. Hover over **Customise** on the top.
2. Click on **Layout**.
3. Move **Year** to the Page panel.
4. Move **Main means of travel to work** and **Area** to Row panel.
5. Keep Column panel empty.
6. Click on **Customise table options**.
7. In **Dimension Member Labels**, tick the **tickbox** of **All Dimensions** and **Use codes**. As the result, the full column shall be ticked.
8. Click on **Customise selection**.
9. Click on **Area** tab.

In **Area** tab:

* 1. Untick both **Total, New Zealand by regional council/area unit** and **Total, New Zealand by territory authority/area unit**.
  2. Click on **Far North District** under **Total, New Zealand by territory authority/area unit**.
  3. On top right corner, click on **Select items**.
  4. Click on **Select level within node**. All Territory Authority levels shall be selected.
  5. Scroll down to the bottom, Untick **Total, Territorial Authority areas** and **Area outside territorial authority**.

1. Click on **Main means of travel to work** tab.

In **Main means of travel to work** tab:

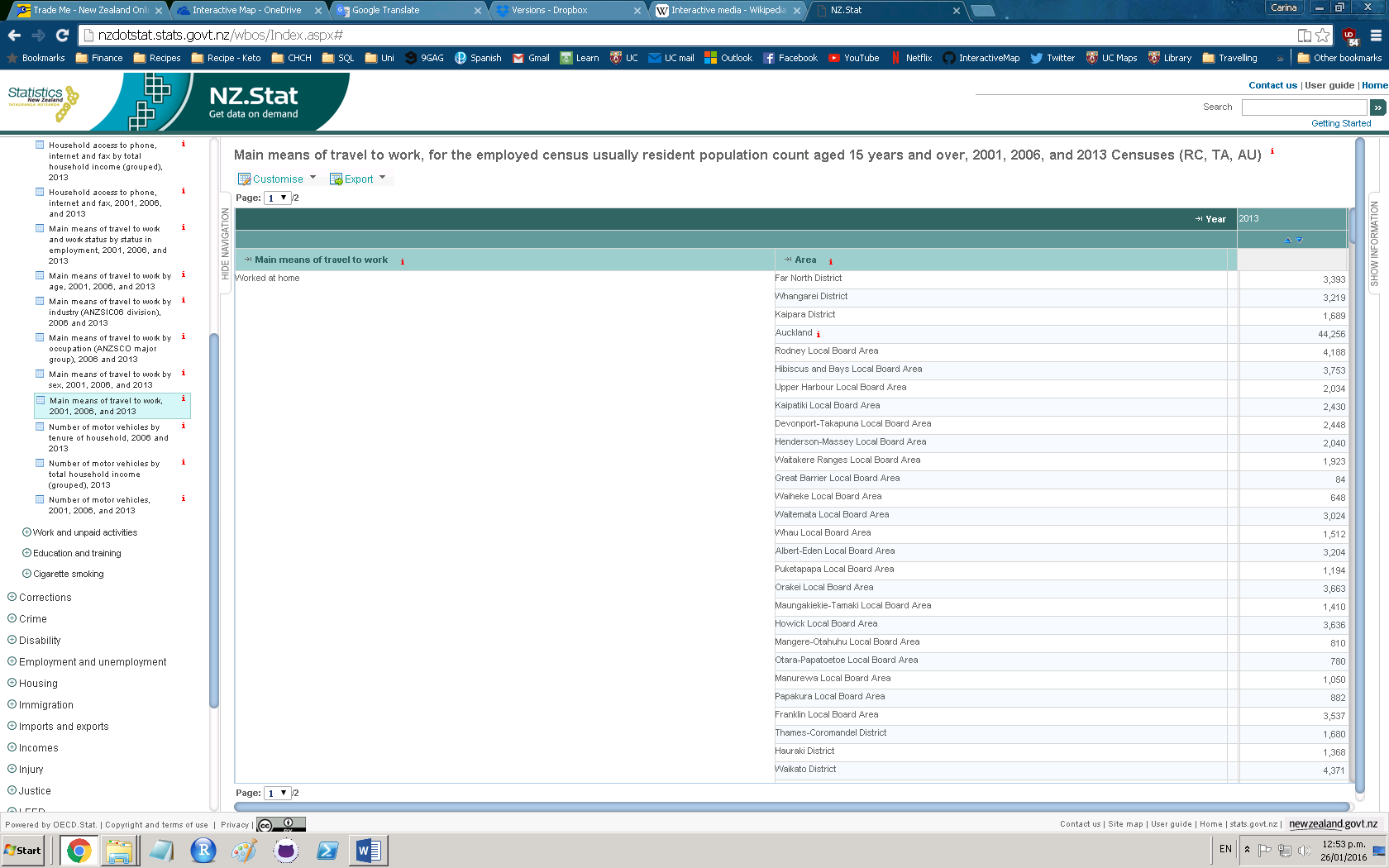
* 1. Untick **Total people, main means of travel to work**.
  2. Click on **Worked at home**.
  3. On top right corner, click on **Select items**.
  4. Click on **Select level within node**. All travel means shall be selected.
  5. Untick **Total People stated** and **Not elsewhere included**.

1. Click on **Year** tab.

In **Year** tab:

* 1. Untick **2001** and **2006**.

1. Click on **View data**.



The page shall display the customized online table (refer to picture above), there might be slight delay depends on the internet speed on the computer that is in used.

To download the online table for further use, do the following:

1. Hover over **Export**.
2. Click on **Excel**.
3. Select the preferred format. The format that is used in this research is Excel, with **Include extra columns for flags** unticked.
4. Click on **Export to XLS file**.

The customized data in selected form shall be downloaded to the computer.

At this point, the spreadsheet that is being used in this research is generated and ready to carry on the rest of the research.

## Functions Selecting Progress

The functions selections progress is separated into two parts: background foundation (FD) and user interface (UI). The background foundation provides and pre-process the data required for the interactivities with the users. And as its name, the users are not able to view the processes. The user interface is the platform that users see, use and obtain information from. The background foundation also receives the indications made by users on user interface, while the user interface sends and displays the generated result.

### **B**ackground foundation

#### Source Data tables

The source data is being stored in the format of Comma Separated Values (csv). Storing and reading data in csv file have the following advantages:

* 1. It can be opened by notepad (of Windows and Linux) and Microsoft Office Excel.
  2. The data fields are separated by comma (",") which makes the data and plain reading more accurate and clear.

There is one key table in this process, it is called geodata. It is the mother table which is used to generate the three subsets: meandata, meanChoices and totalList.

**geodata** reads the data from the spreadsheet downloaded in Primary Data Exploratory section. It contains the codes and descriptions of territory authorities and travel means, as well as the number of people who travel in the selected travel means in the territory authorities.

**meandata** is a subset of geodata but it only contains the codes and descriptions of selected travel means.

**meanChoices** is a subset of meandata, which stores the means codes and names as objects values and names respectively. meanChoices is the options list to travel means checkboxes in User Interface.

**totalList** is a subset of meandata, which contains the total number of people group by territory authorities and the corresponding territory authorities codes. totalList is later used for calculating the percentages of the travel means within territory authorities.

#### Percentage calculation

To colour the maps properly, we require the percentages of the travel means within the same territory authorities to be calculated. Another way to understand the concept, a percentage represents weight of a travel mean being chosen in a territory authority.

The percentages of the travel means are calculated by applying the following the steps to all data rows in the shapefiles data attribute:

* 1. Calculate the total number of people involved in the 2013 census, group by territory authorities (refer to "List of Territory Authorities", Data Classifications in Appendices).
  2. Calculate the percentage of each travel mean within the same territory authority: Divide the number of people using such mean by the sum of people in the corresponding territory authority.

For example, there are 168,219 people from "Christchurch City" (a territory authority) involved in the research. 5,151 people chose "Public bus" (a travel mean) on the survey conducted day. Therefore, the percentage of "Public bus" in "Christchurch City" is calculated as

5,151/168,219 = 0.03062080.

We remain one decimal place in this research because having more decimal places would have the same effect yet it does not have a better reading.

#### Geographic boundary files reading

On Statistics New Zealand website, there are few options available regarding to the geographic boundary files (refer to <http://www.stats.govt.nz/browse_for_stats/Maps_and_geography/Geographic-areas/digital-boundary-files.aspx>). The one we are using here is New Zealand 2013 (NZTM) (246MB), ESRI shapefiles, in Census-based files section.

The shapefiles data is extracted from 2013 Digital Boundaries Generalised Clipped, the ESRI shapefile Output from the ESRI\_Census\_Based\_2013\_NZTM folder.

We then read the shapefiles to the environment by readShapeSpatial function, and remove the territory information where territory that are with names “Chatham Islands Territory” and “Area Outside Territorial Authority” because they are not valid nor containing meaningful data to this research.

Afterwards, we merge the data from geodata to the data attribute of the shapefiles object. Such merging is performed by matching the territory authorities’ codes in both geodata and the shapefile object. Any Not-available (“NA”) data field is replaced by “0” for format purpose. After the merging, data attribute of the shapefiles object shall contain both original shapefiles data and the data from the geodata table.

#### Interval selection

The interval definition requires two components: the interval style and the number of preferred intervals (n).

We provide six options in terms of the class interval styles. These styles are chosen because they are commonly used, well-known and the most representative interval methods to both statisticians and non-statisticians.

The chosen styles are:

*Equal* – it divides the range of the percentages into n parts.

*Fixed breaks* – the data shall be separated by the number of intervals specified by the users.

*Pretty* – it chooses a number of breaks not necessarily equal to n using *pretty*, but likely to be legible.

*Quantiles* – provides quantile breaks.

*Standard Deviation* – it chooses breaks based on *pretty* of the centred and scaled percentages, and may have a number of classes different from n.

The style chosen shall affect the components setup below.

#### dColour definition

We have considered few different options for the colours template.

Initially we tried to create a temporary list and a number of gradient colours (depends on the *n* that is set in *3. Interval selection*). All percentages of the selected travel mean would be gone through and a colour would be assigned to each of them depends on the percentage values. The higher values would have darker colours. The territory authorities’ codes, travel means, the percentages and the assigned colours would be stored in the temporary list. At the end, we colour the New Zealand map by applying the colours assigned in previous steps.

The drawbacks of the initial approach are:

1. All colours to be assigned to each row require calculation or conditional assignment. This would slow down the whole process. Such process would take longer time as *n* increases.
2. It requires more time and effort to edit the code even though the changes to be made are small. This is because editors need to identify which conditional statements they shall make the changes at, therefore, editors shall first read the section of codes.

Overall, the initial approach is good for modelling how the colouring method shall work but not for actual application.

From there, we tried few different methods and functions, we decided to use a combination of colorRampPalette and findColours functions.

colorRampPalette returns functions that interpolate a set of given colors to create new color palettes and color ramps. Referring to the code below, we passes two colours, yellow and red, where yellow stands for the low end of the percentages while red stands for the high end. We use “rgb” here to keep better consistency through the development.

*pal <- colorRampPalette(c("yellow","red"), space= "rgb")*

And then we match the colours to the percentages by using findColours function, passing the interval we gain previously and the colour palette that contains *n* colours.

*len <- length(nclass$brks)*

*colPal <- findColours(nclass, pal(len))*

At this point, all selected percentages of a specific travel mean have matching colours.

#### New Zealand Map (single travel mean)

Because we have decided using the colorRampPalette and findColours functions in *Colour* *definition*, it is easy to plot the map as it can be done by calling plot function.

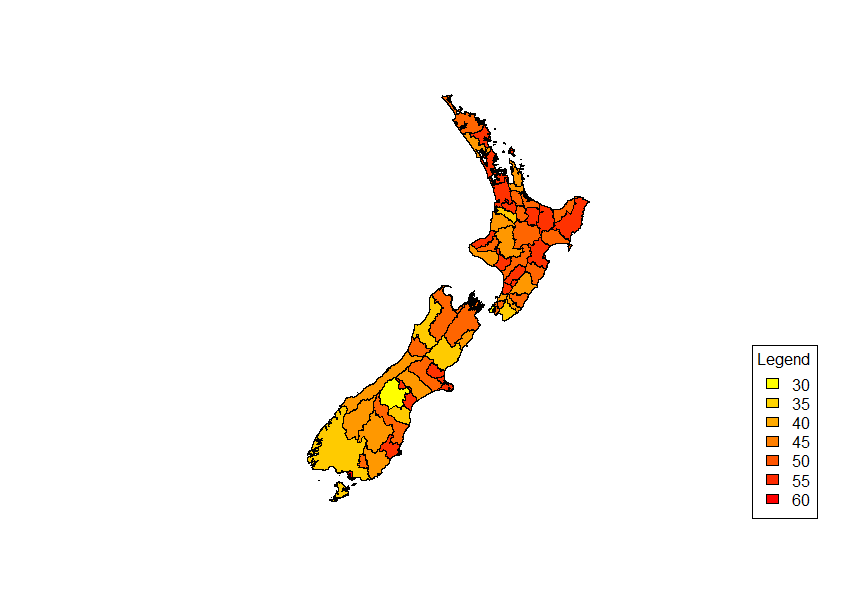
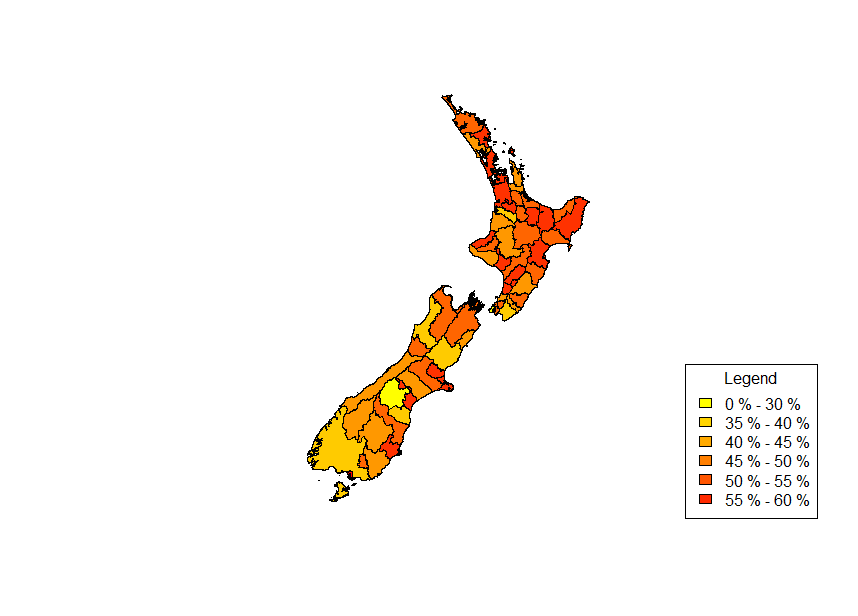
*plot(shape, legend=FALSE, border = "Black", col= colPal)*

In the code above, we call the function plot and set the legend not to be displayed, border colour in map to be black and the colours of the components in map.

The legend text in this map is customised and enhanced:

1. **For all values**: “%” symbol has been added to each percentage. Doing this, it is more obvious to users that the values are in percentages.
2. **First interval**: “0%” has been added to the first interval in order to make the interval clearer.
3. **Intervals between first and last intervals**: Instead of showing either the upper or lower values of the intervals, “-” symbol has been added between the lower and upper values of the corresponding intervals.

Overall, the default legend is: The enhanced legend is:

#### Comparison list for two travel means

The purpose here is to be able to compare two travel means in the same territory authorities and get ready to present the data in the form of coloured map. Colouring the map of two travel means requires calculating the corresponding colours by their percentages and storing them clearly.

We first tried to build a two-way table to meet the requirements. However, most of the two-way tables are contingency tables which counts the number of occurrences that meet both row and column criteria. Manipulating the values in the data fields is possible but it requires longer research time, therefore it cannot be accomplished.

Our second attempt was building a matrix and fitting the values by the ranks of the territory authorities in both travel means. One of the difficulties that occurred is manipulating. Given the structure of matrix, we needed to specify the row and column positions when inserting the values. This involves an unnecessary amount of symbols and specifications.

Exploring few other methods, we have decided to compose the desiring list by following the logic:

1. Generate a subset list named **listx** of shapefiles data attribute, selecting the rows which travel means are the same as the first user-selected travel mean.
2. Generate another subset list named **listy** of shapefiles data attribute, selecting the rows which travel means are the same as the second user-selected travel mean.
3. Order both lists by percentages in ascending order.
4. Create new columns in both lists named xpos and ypos respectively, which stores the ranks of the data rows in separated lists.
5. Merge listy into listx by matching their territory authorities’ codes.
6. Calculate the corresponding colours (in RGB) using xpos and ypos values, store the colours in a new column named mix.

Below is the full code for the function:

#### New Zealand map (of two travel means)

After building the list in *7. Comparison list for two travel means*, we now have a list of calculated colours for the selected travel means. At this point, colouring the map of two travel means is an easy task. All we need to do is calling the plot function, passing the mix column in the list as the colour palette.

### User interface

Shiny is an open source R package. Developers can build web applications by using R along with it as the web framework. It can also be used with HTML, CSS, and JavaScript for better styling and enhanced functionalities. More information can be found at <http://shiny.rstudio.com/>.

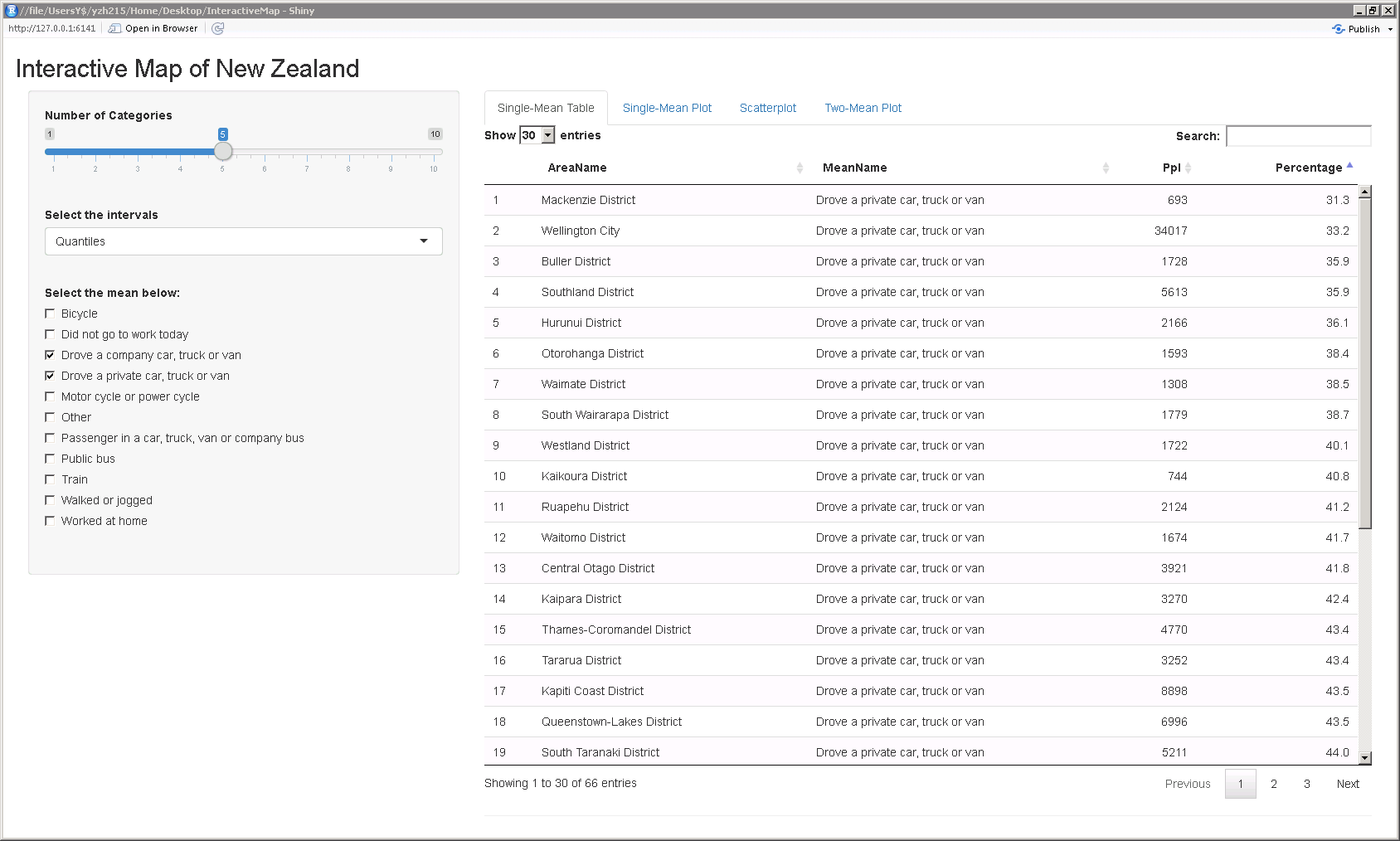
Shiny plays a major and important role in building the user interface in this research. We have the following main components in the code:

#### Interval definition

To define the interval, it requires two parameters to be set by users. These parameters are Number of Categories and Interval style.

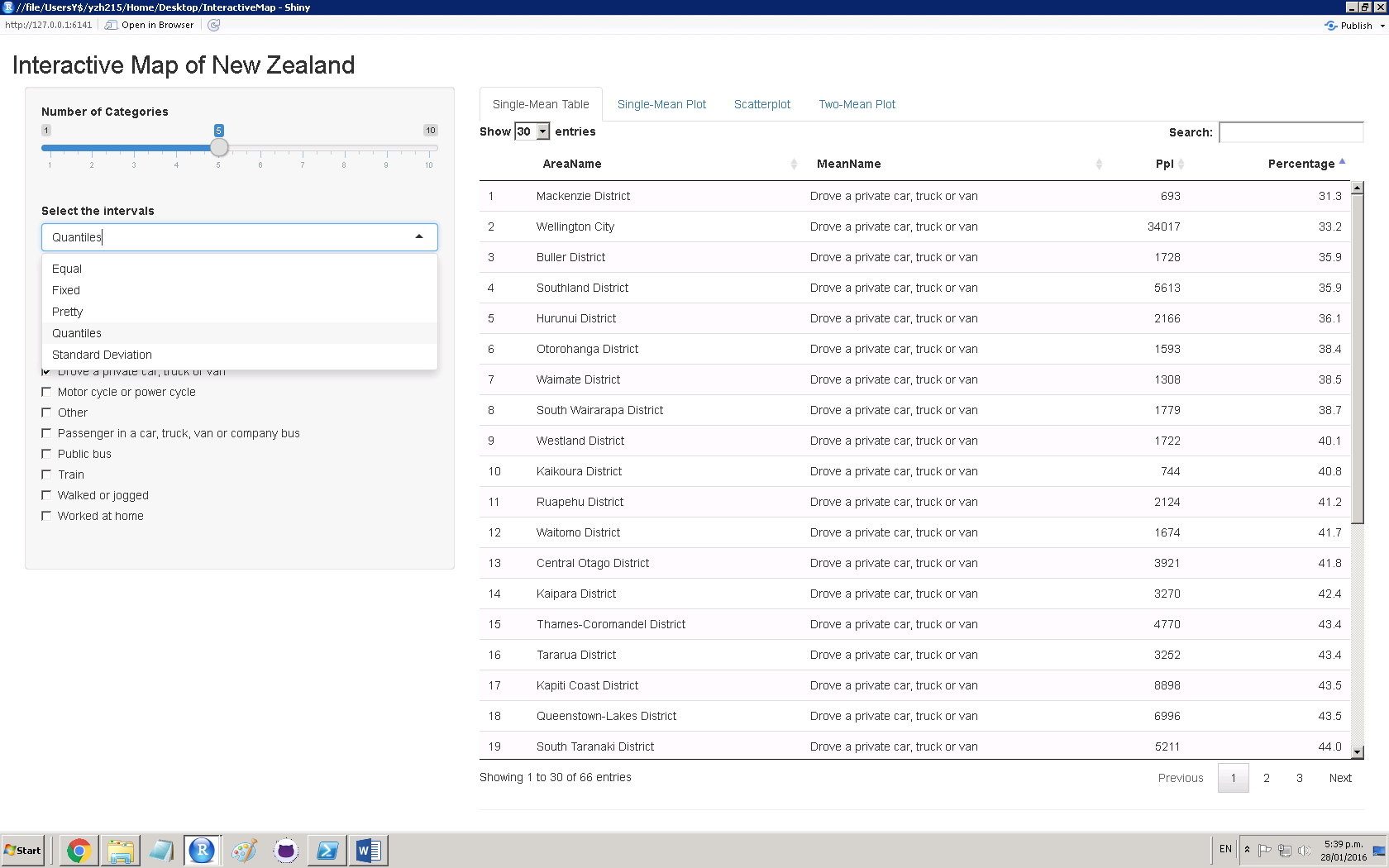
* 1. Number of Categories

Because users want to have different numbers of intervals when they look at the percentages and the map, we use a slidebar to obtain the parameter. Comparing to textbox where users can enter anything (numeric, character and symbol etc.), we can make sure the users are dividing the percentages within a reasonable range of number, without returning any warning when users enter invalid information in the textbox. Overall, we have more data validity control on the parameter yet limit the usage.



* 1. Interval style

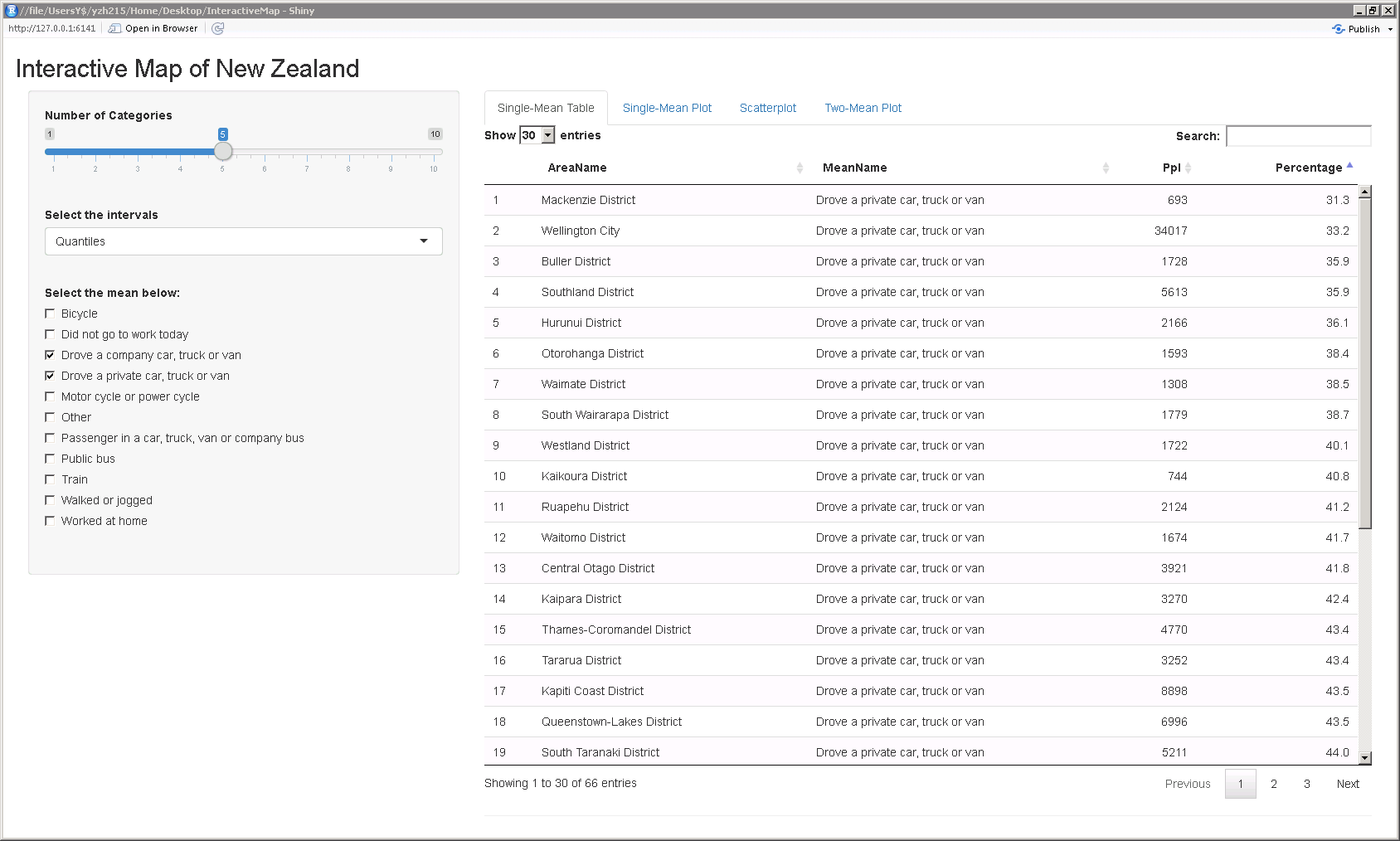
We use the dropdown box for the interval style parameter. This is because we provide only those which are commonly known and useful to all users to avoid any confusion and misunderstanding. Therefore, selecting the style from the list is the best solution in this research.



#### Travel means

Users select one travel mean to view the relevant data and the coloured map of the travel mean. We allow users select two travel means at the most under the circumstance when they want to compare usages of two travel means in territory authorities.

In this case, we choose a group of checkboxes for this parameter. Group of checkboxes enable users view all options at once and clearly see which option(s) they have chosen. The chosen travel means are passed to the background function for updating and displaying result purposes.



## Application Discussion

There are four tabs in the web application. These four tabs have been separated into two sets: one travel mean and two travel means. Each set contains two tabs: the table tab and the plot tab. We are looking at these tabs in details.

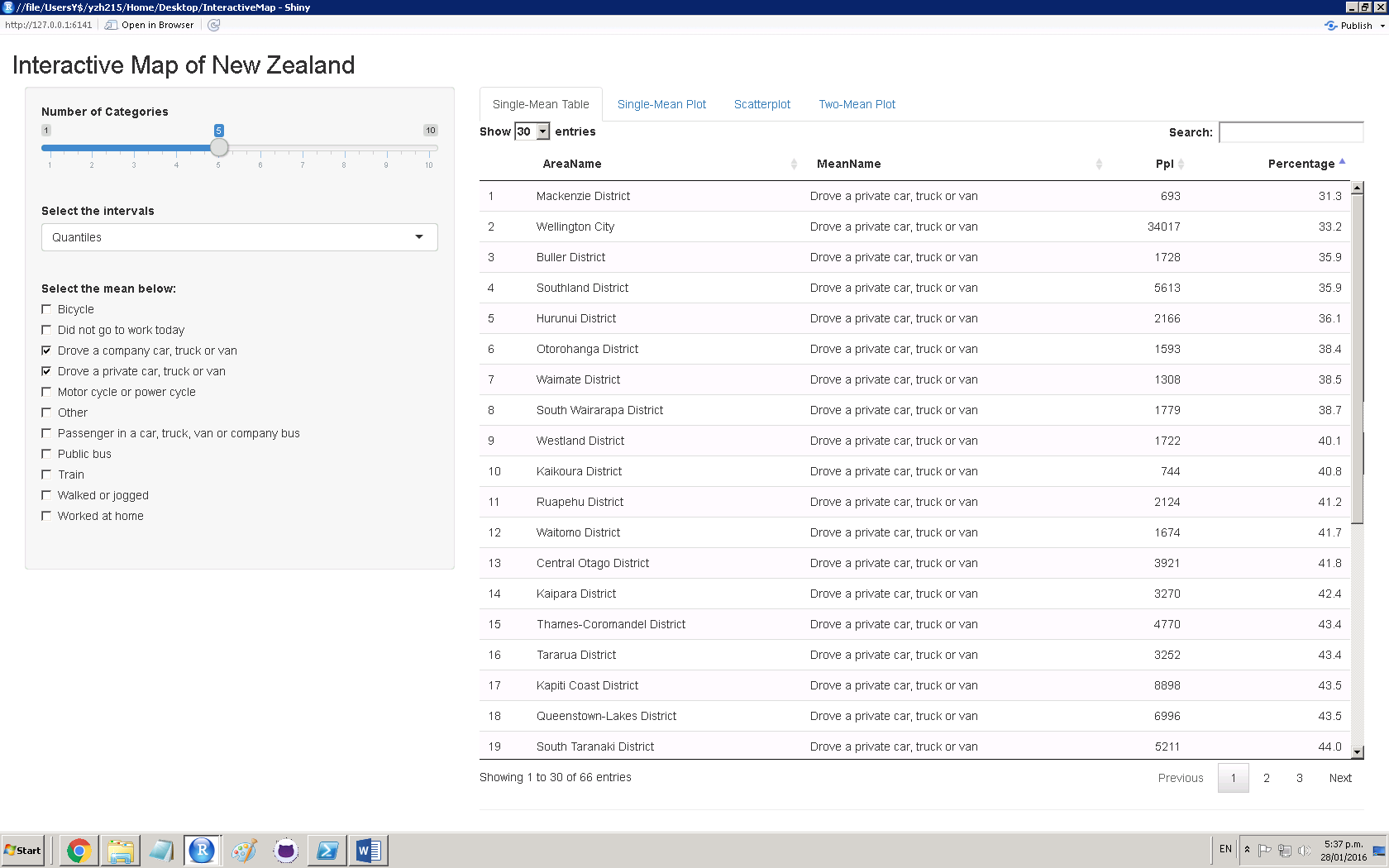
#### Data Table of the selected travel mean

A data table shall be displayed here. It aims to enable users to have an overview of the travel mean in all territory authorities. The data table is using the subset of the data attribute of the shapefiles object.

If users select one travel mean, the table selects and displays all data rows that are of the travel mean. Each of these data rows contains: territory authorities names, travel mean name, number of people chose such travel mean, the percentages of the travel mean in corresponding territory authorities.

If users select two travel means, the table displays the relevant information of the last selected travel mean.

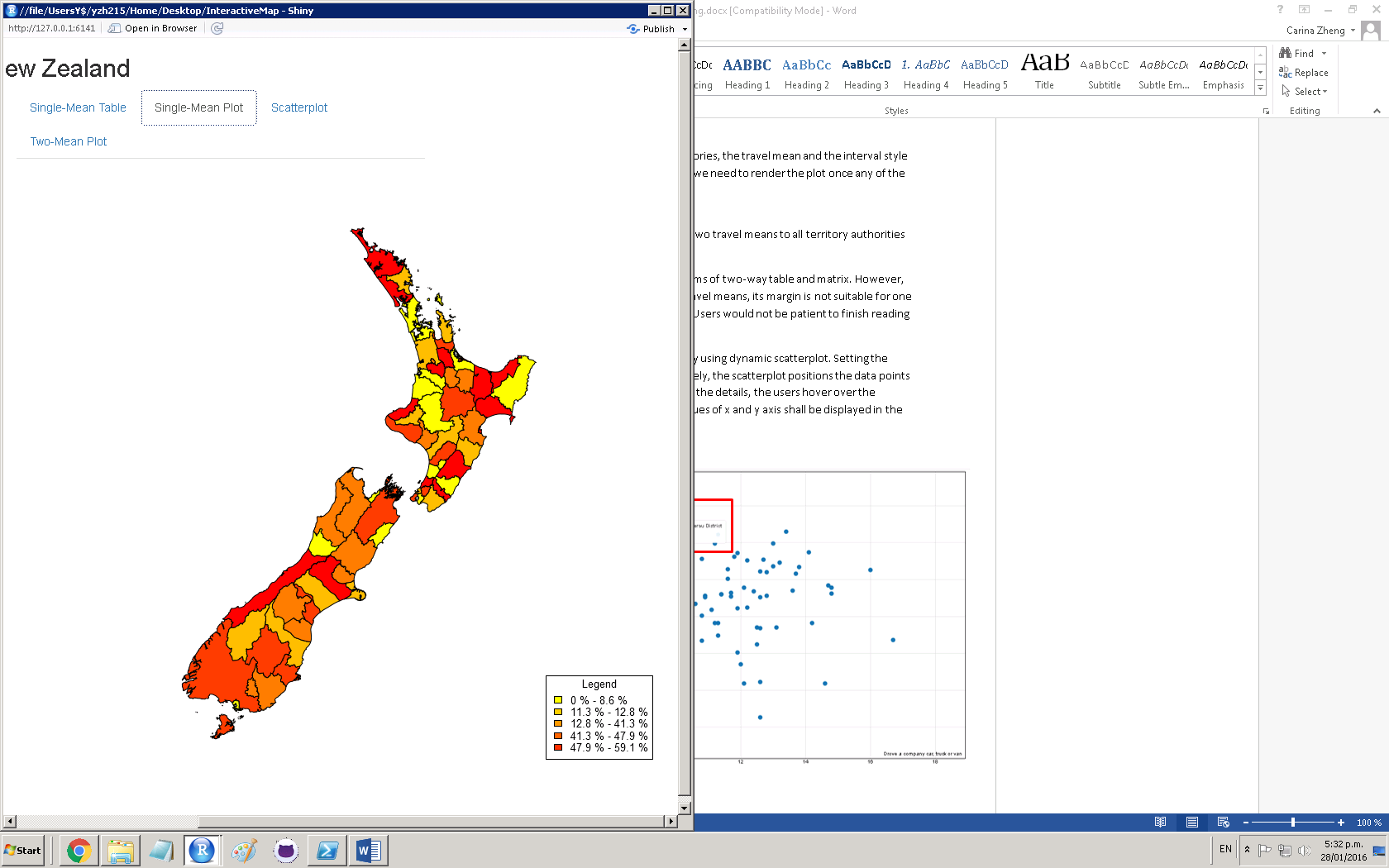
One of the key properties is the data table shall be rendered each time the travel mean changes. This is of the purpose of updating the table so that the system is aware of the change of parameter in order to update table content.



#### New Zealand Map of single travel mean

The aim of this component is to display the map which is coloured by the percentages of the rows. It represents the percentages by the intensity of the colours. The colours are changing from yellow to red, where yellow represents the lower end of the percentages and red represents the higher end of the percentages. The territory authorities where have orange colour stands for the middle range of the percentages.

To draw the map, we need to pass number of categories, the travel mean and the interval style parameters to draw the map. For the same reason, we need to render the plot once any of the parameter is updated.

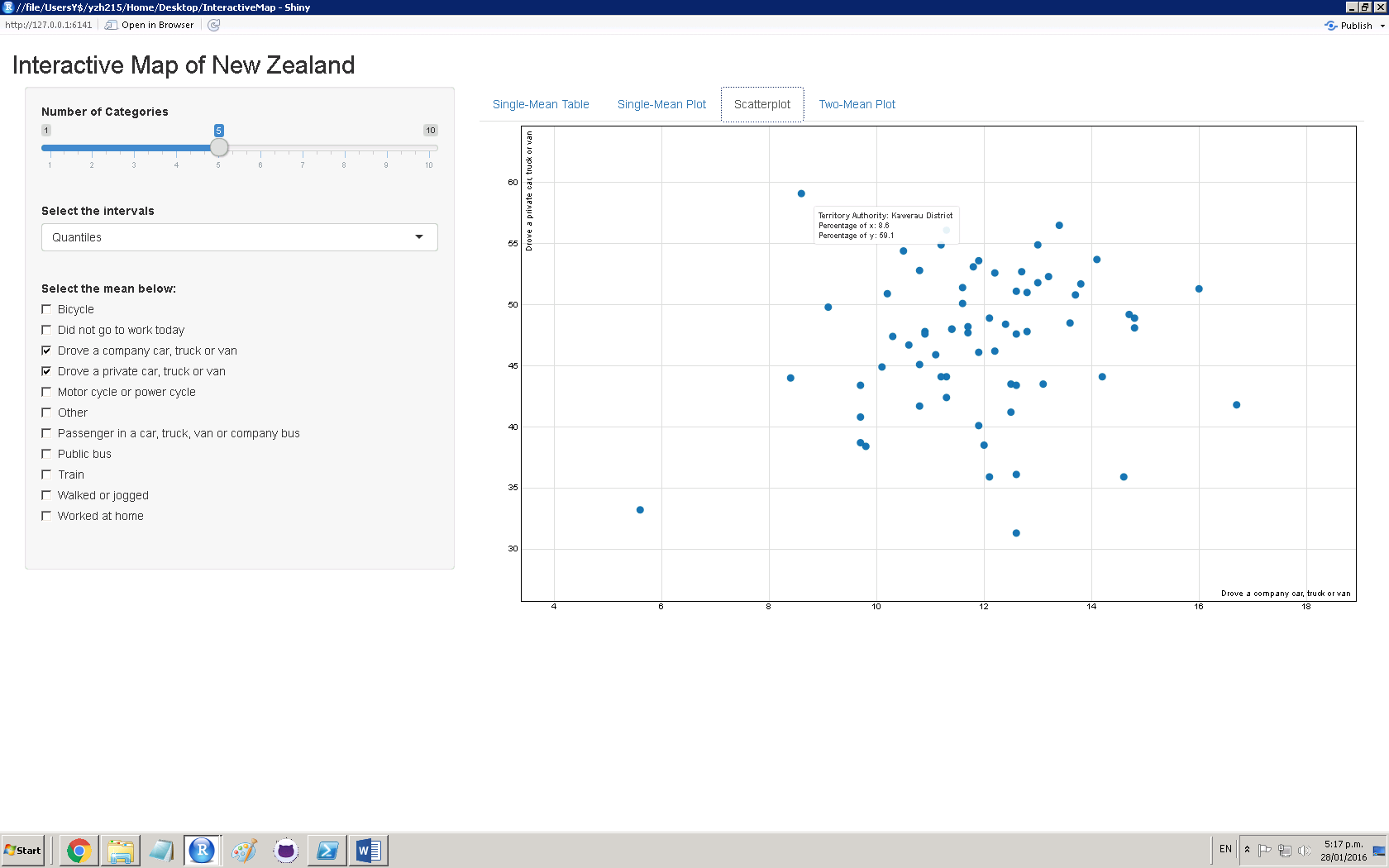


#### Dynamic Scatterplot

The scatterplot is aim to display the information of two travel means to all territory authorities in a comparing manner.

In our first attempt, we format the data into the forms of two-way table and matrix. However, given the amount of data to be displayed for two travel means, its margin is not suitable for one webpage and leads to troubles in reading the data. Users would not be patient to finish reading the full information given its downsides.

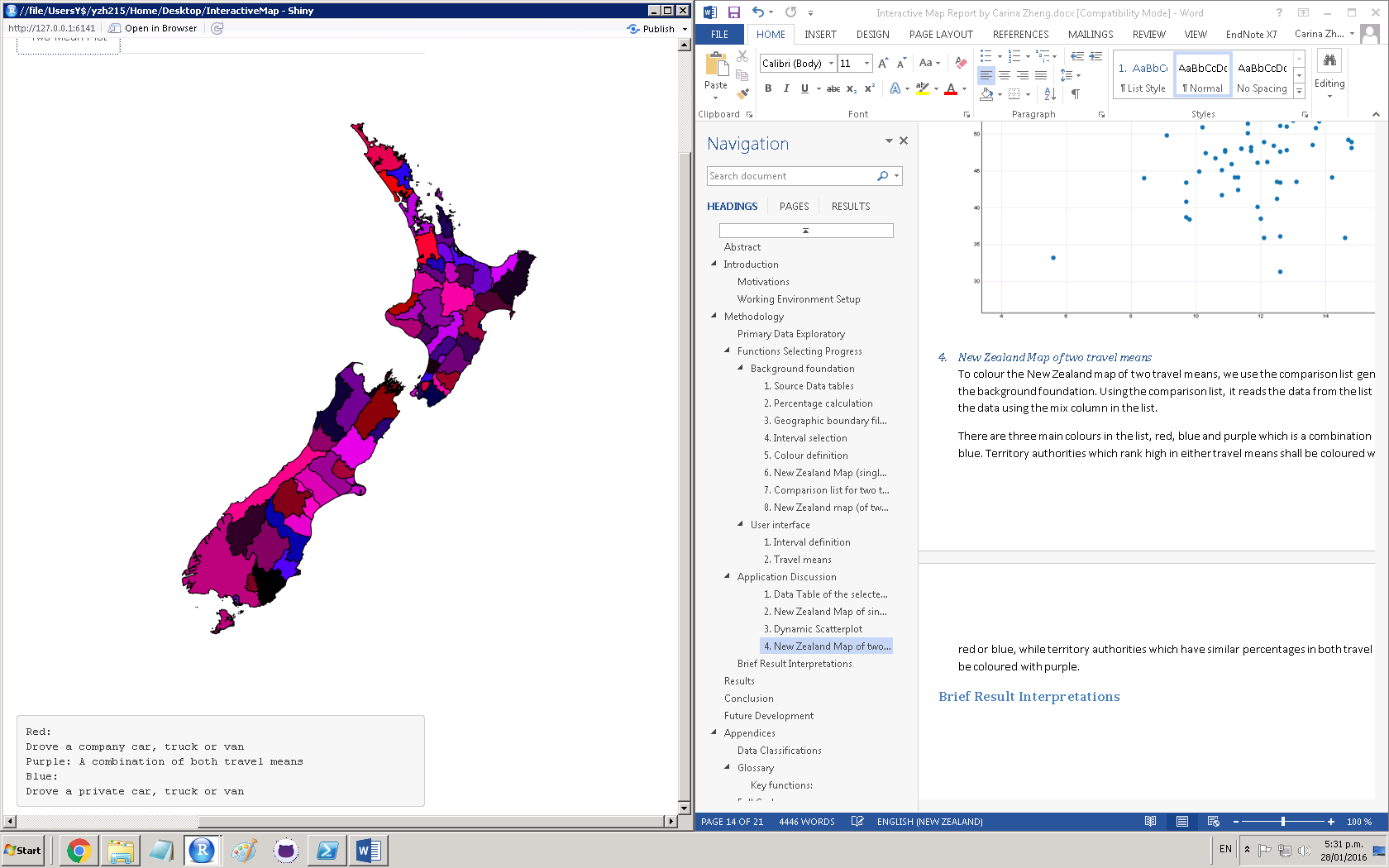
After that, we have decided to represent the data by using dynamic scatterplot. Setting the percentages in both means as x and y axis respectively, the scatterplot positions the data points by the data’s percentages in both means. To display the details, the users hover over the desiring data point and the territory authorities, values of x and y axis shall be displayed in the dialog box next to the data point.



#### New Zealand Map of two travel means

To colour the New Zealand map of two travel means, we use the comparison list generated in the background foundation. Using the comparison list, it reads the data from the list and colour the data using the mix column in the list.

There are three main colours in the list, red, blue and purple which is a combination of red and blue. Territory authorities which rank high in either travel means shall be coloured with intensive red or blue, while territory authorities which have similar percentages in both travel means shall be coloured with purple.



# Results

This is where I do the result interpretations…

# Conclusion

1. The components of a good interactive app/site are….
2. The travel means ….

# Future Development

Given the short time frame of the research project, some ideas cannot be performed. However, they shall be descripted and discussed in this section.

1. Interactive Linear Regression Plot.

Summary:

1. What I haven’t done/gone through because of I’m short of time
2. What I would do if I got more time
3. Suggestions and thoughts to data visualization and interactive map

# Appendices

## Data Classifications

1. List of Territory Authorities:
   * 001 Far North District
   * 002 Whangarei District
   * 003 Kaipara District
   * 011 Thames-Coromandel District
   * 012 Hauraki District
   * 013 Waikato District
   * 015 Matamata-Piako District
   * 016 Hamilton City
   * 017 Waipa District
   * 018 Otorohanga District
   * 019 South Waikato District
   * 020 Waitomo District
   * 021 Taupo District
   * 022 Western Bay of Plenty District
   * 023 Tauranga City
   * 024 Rotorua District
   * 025 Whakatane District
   * 026 Kawerau District
   * 027 Opotiki District
   * 028 Gisborne District
   * 029 Wairoa District
   * 030 Hastings District
   * 031 Napier City
   * 032 Central Hawke's Bay District
   * 033 New Plymouth District
   * 034 Stratford District
   * 035 South Taranaki District
   * 036 Ruapehu District
   * 037 Wanganui District
   * 038 Rangitikei District
   * 039 Manawatu District
   * 040 Palmerston North City
   * 041 Tararua District
   * 042 Horowhenua District
   * 043 Kapiti Coast District
   * 044 Porirua City
   * 045 Upper Hutt City
   * 046 Lower Hutt City
   * 047 Wellington City
   * 048 Masterton District
   * 049 Carterton District
   * 050 South Wairarapa District
   * 051 Tasman District
   * 052 Nelson City
   * 053 Marlborough District
   * 054 Kaikoura District
   * 055 Buller District
   * 056 Grey District
   * 057 Westland District
   * 058 Hurunui District
   * 059 Waimakariri District
   * 060 Christchurch City
   * 062 Selwyn District
   * 063 Ashburton District
   * 064 Timaru District
   * 065 Mackenzie District
   * 066 Waimate District
   * 068 Waitaki District
   * 069 Central Otago District
   * 070 Queenstown-Lakes District
   * 071 Dunedin City
   * 072 Clutha District
   * 073 Southland District
   * 074 Gore District
   * 075 Invercargill City
   * 076 Auckland
2. Main means of travel to work:
   * 01 Worked at home
   * 02 Did not go to work today
   * 03 Drove a private car, truck or van
   * 04 Drove a company car, truck or van
   * 05 Passenger in a car, truck, van or company bus
   * 06 Public bus
   * 07 Train
   * 08 Motor cycle or power cycle
   * 09 Bicycle
   * 10 Walked or jogged
   * 15 Other (includes taxi, ferry, helicopter, aeroplane)

## Glossary

### Key functions:

1. require("package\_name") || install.packages("package\_name ")

Install the named package if it is not install in the environment.

1. setwd("folder\_path")

Set the home directory of the environment.

1. geodata <- read.csv('geodata.csv'
2. , col.names= c('AreaCode','AreaName','AreaFull','MeanCode','MeanName','MeanFull','Ppl')
3. , header= FALSE
4. , sep = ','
5. , numerals = c('no.loss'))meandata <- unique(geodata[c('MeanCode', 'MeanName', 'MeanFull') ] )

Rea the csv file into the environment.

1. meandata <- meandata[order(meandata$MeanName),]

Order the data in meandata list by MeanName variable.

1. meanChoices <- as.character(meandata$MeanCode)

Convert the MeanCode variable to from type of factor to type of character.

## Full Code

# References

1. Ward, Grinstein and Keim, 2010
2. Statistics New Zealand (2015) <http://stats.govt.nz/> (checked on 22nd December 2015)

# Acknowledgement

This research was supported by Statistics New Zealand. We thank staff from Statistics New Zealand and University of Canterbury who provided professional advices that greatly assisted the research.

We would also love to specially thank Richard Penny, supervisor of the research, for sharing his expertise and his patience to any question raised.

**LOG: LAST UPDATED ON 28/1/16 18.08**