**A puzzle**

In my [last blog post](http://ellisp.github.io/blog/2018/01/23/recruiting), I gave a hypothetical test exercise for candidates in a recruitment process, involving the International Visitor Survey (IVS) from the Ministry of Business, Innovation and Employment (MBIE). I once managed the team responsible for that survey and other sector data, so I was familiar with it, knew the exercise made sense, and knew the microdata are publicly available (albeit not documented as well as anyone would like – a known issue, it’s never possible to do everything). In [follow-up discussion on Facebook](https://www.facebook.com/peterstats/) I said I’d never give an exercise that I couldn’t complete myself in a reasonable amount of time, so I thought a good follow-up post would be to show the sort of thing I’d expected from candidates.

[[](https://en.wikipedia.org/wiki/Queenstown,_New_Zealand)](https://en.wikipedia.org/wiki/Queenstown,_New_Zealand)

Well, this particular exercise failed to pass that test because it took me too long to do it myself. Partly because there were too many things mentioned (comparing visitors to Queenstown to others; spend; activities; length of stay; origin); and partly for the more interesting reason that there was a real fish-hook hidden in the data. Rather than boringly talk about how the recruitment test needs to be cut down a bit, this blog is about that fish-hook, which involves the perils of analysis of data where concepts have a one-to-many relationship, which in the right circumstances can make it look like:

Everyone is above average here!

Before I dive into it, here’s the problem at a high level. One of the things that needed to be done to answer the questions posed in exercise was the average spend of tourists who stayed overnight in Queenstown (a town in New Zealand’s south island, heavily touristed, ideal for outdoors activity, extremely beautiful – see picture above) compared to those who didn’t. The obvious chart to address this is this one:

MBIE processes and publishes the IVS data on a quarterly basis and there’s quite a seasonal effect even in average spend, so the thicker, darker lines are seasonally adjusted whereas the thin lines are the original.

So far so good; the graph shows nicely that people visiting Queenstown, many of them skiers, and nearly all of them paying relatively high accommodation and hospitality prices, spend more money on their trip to New Zealand than do other visitors. The hitch came in when I tried running the same code with other destinations, such as Christchurch, Dunedin, Auckland, Wellington, Gisborne, Totaranui, Golden Bay… nearly everywhere I tried, the graphic looked similar to this. Tourists who stay at least one night in region X spend more than do tourists who don’t, for a very broad collection of X.

Let’s see how that can be the case.

**Loading up the IVS**

First, let’s set up our R session and download the IVS data. The data comes in a zip file from the MBIE website, which contains 26 different CSV files. Each of these files is actually a copy of a view in MBIE’s own database, so by doing analysis with these files we should be able to replicate exactly the approach used in MBIE itself. We won’t need all 26 views, but it’s just as easy to load them all in at once. I considered putting them into a database, to even closer replicate the in-MBIE experience, but the total size of the data is only about 230MB (when unzipped) and it doesn’t take long to load them all into memory, so I went with the simpler approach of keeping everything in R.

library(tidyverse)

library(data.table)

library(scales)

library(stringr)

library(forcats)

library(ggseas)

library(testthat)

#========Download and import data===============

# download survey data from the MBIE (Ministry of Business, Innovation and Employment) website

download.file("http://www.mbie.govt.nz/info-services/sectors-industries/tourism/tourism-research-data/ivs/documents-image-library/vw\_IVS.zip",

mode = "wb", destfile = "vw\_IVS.zip")

unzip("vw\_IVS.zip")

# list of CSV files, each one of which is a copy of a view in the MBIE database:

files <- list.files("IVS")

# import each CSV file as a data.table/data.frame object in RAM with the name of the view

for(i in 1:length(files)){

print(paste("Importing", files[i]))

tmp <- fread(paste0("IVS/", files[i]))

# knock off the ".csv" at the end of each file name, for the name of the data frane in R:

dfname <- str\_sub(files[i], end = -5)

assign(dfname, tmp)

}

The most important table is the one called vw\_IVSSurveyMainHeader. This has all the survey responses (and artificial variables, like sampling cluster and population weight) that make sense to store on a table of one row per respondent. So observations for which there is only one of for each person – like country of residence next year, purpose of visit, total spend in New Zealand, age and sex – are all stored here. A lot of analysis can be done with just this table. For example, here’s a chart of quarterly spend and visitor numbers by purpose of visit:

The seasonal adjustment here (the thick line) is done on the fly in the ggplot2 pipeline, using X13-SEATS-ARIMA via the seasonal and ggseas R packages, with the line beginning stat\_seas in the code below. In this graphic, notice the way that numbers of holiday makers and (even more so) their spend per visitor took off in the boom beginning around 2014. The numbers of visitors actually come from the arrival/departure cards and get reflected in the IVS via its weighting, so there should be no likelihood of error in the right hand facet. I don’t know what’s behind the jump in “VFR” (visiting friends and relatives) visitor numbers in the early 2000s.

Code for that chart:

#=============Familiarisation============

# First analysis is just to get a feel for the data; not used in the final report

head(vw\_IVSSurveyMainHeader)

# total spend by purpose of visit

# Note that "WeightedSpend" just means outlier-treated; you need to

# use PopulationWeight to get the actual survey weights

spend\_pov <- vw\_IVSSurveyMainHeader %>%

group\_by(Qtr, POV) %>%

summarise(total\_spend = sum(WeightedSpend \* PopulationWeight),

visitors = sum(PopulationWeight),

sample\_size = n())

# graphic

spend\_pov %>%

# convert the "1997 1" format into a number:

mutate(qtr = as.numeric(str\_sub(Qtr, start = -1)),

yr = as.numeric(str\_sub(Qtr, end = 4)),

yr\_qtr = yr + (qtr - 0.5) / 4) %>%

# reshape:

ungroup() %>%

select(yr\_qtr, POV, total\_spend, visitors) %>%

gather(variable, value, -yr\_qtr, -POV) %>%

mutate(POV = fct\_reorder(POV, -value)) %>%

ggplot(aes(x = yr\_qtr, y = value, colour = POV)) +

facet\_wrap(~variable, scale = "free\_y") +

geom\_line(alpha = 0.3) +

stat\_seas(start = c(1997, 1), size = 1.2) +

scale\_y\_continuous("", label = comma) +

ggtitle("Weighted spend and visitor numbers, visits to New Zealand",

"Highly seasonal quarterly data") +

labs(x = "", colour = "Purpose\nof Visit",

caption = "Source: MBIE International Visitor Survey")

**Checking out spend and activities for visitors to one area**

Next step is to identify which tourists included an overnight stay in Queenstown in their visit, and compare their average spend, activities and length of stay to those who didn’t. In addition to the plot that I used near the beginning of the post, this gets us this one:

We can see that like spend, the Queenstown visitors stay longer in New Zealand, and do more activities while in New Zealand, than do visitors who did not overnight in Queenstown. Note the very obvious break in the “activities” series in 2013. This came with a major redevelopment of the survey, moving most of the data collection on line and rationalising the questionnaire for shorter interviews. While spend and length of stay are pretty much comparable (and in fact, great lengths were gone to to backcast the old spend data onto the same basis as the new survey), this wasn’t possible with all the secondary characteristics such as activities, transport types, etc. This break in series is responsible for other quirks too, such as the possibility of respondents reporting no locations overnighted in, which only appears in 2013; another artefact of how the old questionnaire and data model was translated into the new in the best way that could be managed.

A few points to note in the analysis below, creating these charts:

* Visitors are asked (from 2013 onwards) what activities they did in their *whole New Zealand trip*, not broken down by region. The historical data has been put in the same data model, but the change in mode and questionnaire still leads to a series break. As visitors have multiple activities per trip to New Zealand, the activities-respondents combinations are stored in their own long, thin vw\_IVSActivities table.
* Visitors can stay overnight in zero or more locations (some people fly in and out on the same day), so similarly the locations-respondents combinations are stored in their own table, vw\_IVSItineraryPlaces. There are 514 point locations respondents are allowed to choose from (using an on-screen map) as overnighting locations.
* I defined my own palette, with colours named “Visited Queenstown”, “Did not visit Queenstown” etc, to use with scale\_colour\_manual() in order to ensure colours were used consistently in all charts in the analysis.
* Like activities, spend and length of stay are only analysed here on the basis of overall spend and length of stay in *New Zealand*, not just the Queenstown leg. Spend data is not available in this survey for Queenstown (other official statistics, based on electronic card data, can give spend data to that level, but not allocated to individual respondents); length of stay in all of New Zealand was part of the original research question.
* I separated the average spend chart from the faceted plot showing activities and length of stay, partly so spend could have a $ sign in its scale.

#================Activities=================

head(vw\_IVSActivities)

# hmm, an interesting set of "Activities" but we have no way of identifying which are "good" or not

# so we will just count them per person/visit

# number of activities per person

act\_pp <- vw\_IVSActivities %>%

group\_by(SurveyResponseID) %>%

summarise(number\_activities = n())

#================Queenstown=====================

head(vw\_IVSItineraryPlaces)

# 514 distinct places listed:

sort(unique(vw\_IVSItineraryPlaces$WhereStayed))

places\_visited <- vw\_IVSItineraryPlaces %>%

group\_by(SurveyResponseID) %>%

summarise(number\_places\_visited = length(unique(WhereStayed)),

visited\_queenstown = ifelse("Queenstown" %in% unique(WhereStayed),

"Visited Queenstown", "Did not visit Queenstown"))

qt\_visitors <- places\_visited %>%

right\_join(vw\_IVSSurveyMainHeader, by = "SurveyResponseID") %>%

mutate(qtr = as.numeric(str\_sub(Qtr, start = -1)),

yr = as.numeric(str\_sub(Qtr, end = 4)),

yr\_qtr = yr + (qtr - 0.5) / 4,

visited\_queenstown = ifelse(is.na(visited\_queenstown), "Did not visit \*anywhere\*", visited\_queenstown)) %>%

left\_join(act\_pp, by = "SurveyResponseID") %>%

group\_by(visited\_queenstown, yr\_qtr)

expect\_equal(nrow(qt\_visitors), nrow(vw\_IVSSurveyMainHeader))

# Two variables look to be about length of stay; but NoDaysInNZ has 120,000+ NA values so we will use the other

# option which has only 196

summary(vw\_IVSSurveyMainHeader[ , c("NoDaysInNZ", "LengthOfStay")])

qt\_visitors\_sum <- qt\_visitors %>%

summarise(total\_spend = sum(WeightedSpend \* PopulationWeight),

total\_days = sum(LengthOfStay \* PopulationWeight, na.rm = TRUE),

total\_activities = sum(number\_activities \* PopulationWeight, na.rm = TRUE),

visitors = sum(PopulationWeight),

spend\_per\_visitor = total\_spend / visitors,

days\_per\_visitor = total\_days / visitors,

activities\_per\_visitor = total\_activities / visitors,

sample\_size = n()) %>%

ungroup() %>%

mutate(visited\_queenstown = fct\_reorder(visited\_queenstown, total\_spend))

palette <- c("Visited Queenstown" = "red", "Did not visit Queenstown" = "blue", "Did not visit \*anywhere\*" = "grey")

# do they spend more per visit?

qt\_visitors\_sum %>%

ggplot(aes(x = yr\_qtr, y = spend\_per\_visitor, colour = visited\_queenstown)) +

geom\_line(alpha = 0.15) +

stat\_stl(s.window = 7, size = 1.2) +

theme(legend.position = "right") +

scale\_colour\_manual("", values = palette) +

scale\_y\_continuous("Spend per visitor\n(seasonally adjusted)", label = dollar) +

labs(x = "", caption = "Source: MBIE International Visitor Survey") +

ggtitle("Total spend while in New Zealand",

"Comparing visitors who went to Queenstown with those who did not")

# do they stay more days and do more activities?

qt\_visitors\_sum %>%

select(visited\_queenstown, yr\_qtr, days\_per\_visitor, activities\_per\_visitor) %>%

gather(variable, value, -visited\_queenstown, -yr\_qtr) %>%

ggplot(aes(x = yr\_qtr, y = value, colour = visited\_queenstown)) +

facet\_wrap(~variable, scales = "free\_y") +

geom\_line(alpha = 0.15) +

stat\_stl(s.window = 7, size = 1.2) +

scale\_colour\_manual("", values = palette) +

scale\_y\_continuous("Value\n(seasonally adjusted)", label = comma) +

labs(x = "", caption = "Source: MBIE International Visitor Survey") +

ggtitle("Activities and length of stay while in New Zealand",

"Comparing visitors who went to Queenstown with those who did not")

**Let’s look at all locations at once…**

Now is where it gets puzzling. Running the code above but substituting nearly any other location for Queenstown gets a similar pattern – higher average spend, length of stay, and number of activities for the people who visited that location than those who didn’t. Let’s go straight to a chart of the top 150 locations – not all of them labelled – showing the average spend of people who overnighted in that location:

We can see that a big majority of places have above-average spend, shown by the vertical blue line at about $3,200. The same applies to the other value variables we’ve considered (number of activities, and length of stay).

The code that produces this is shown below. An important thing to note here is that many individuals will be counted against multiple locations. There’s nothing wrong with this, it’s by design, but it’s important to note! Individuals who stayed overnight at more locations get counted more than those who stayed at fewer locations – something that gives us the clue for how all locations can be “above average”.

#===============detective work into how all locations seem to have higher spenders on average=========================

# each distinct place visited at least once by each person:

ip <- vw\_IVSItineraryPlaces %>%

select(SurveyResponseID, WhereStayed) %>%

distinct()

# total and average spend for each location, 2014 onwards:

all\_locs <- vw\_IVSSurveyMainHeader %>%

filter(Year >= 2014) %>%

left\_join(ip, by = "SurveyResponseID") %>%

mutate(WhereStayed = ifelse(is.na(WhereStayed), "Did not stay overnight at any location", WhereStayed)) %>%

group\_by(WhereStayed) %>%

summarise(total\_spend = sum(WeightedSpend \* PopulationWeight, na.rm = TRUE),

visitors = sum(PopulationWeight, na.rm = TRUE),

mean\_spend = total\_spend / visitors) %>%

filter(visitors > 0) %>%

arrange(desc(mean\_spend))

# overall average for comparison

overall\_mean <- vw\_IVSSurveyMainHeader %>%

filter(Year >= 2014) %>%

summarise(mean\_spend = sum(WeightedSpend \* PopulationWeight) / sum(PopulationWeight))

# graphic of just the 150 biggest

all\_locs %>%

filter(rank(-visitors) <= 150) %>%

mutate(label = ifelse(rank(-visitors) < 20, WhereStayed, "")) %>%

mutate(WhereStayed = fct\_reorder(WhereStayed, mean\_spend)) %>%

ggplot(aes(x = mean\_spend, y = WhereStayed)) +

geom\_vline(xintercept = overall\_mean$mean\_spend, colour = "steelblue") +

geom\_point(aes(size = visitors)) +

geom\_text(aes(label = label), hjust = 0, nudge\_x = 100, size = 2) +

scale\_x\_continuous("Mean spend per visit from 2014 to 2017", label = dollar) +

scale\_size\_area("Number of visitors\nsince 2014", label = comma) +

labs(caption = "Source: MBIE International Visitor Survey") +

ggtitle("Mean spend of visitors to the 150 most visited locations",

"Only the top 20 locations are labelled") +

theme(axis.text.y = element\_blank(),

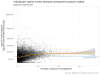
legend.position = "right")

**Relationship between visits and spend**

So the mystery becomes clear once we think through the issue:

“if individuals who overnight at more locations get counted more times, is there any relationship between number of locations visited and individuals’ spend?”

… and of course there is such a relationship. Here it is at the individual level:



Generally speaking, the more locations visited overnight (at least up to about 25), the higher individuals’ total spend in New Zealand. In fact, this makes perfect sense once you think about it.

[BTW, the image above is in PNG format rather than the SVG I usually use, to save space. As there is a semi-transparent dot for every respondent of the survey since 2014, that’s 30,000+ points for the SVG to hold information on and individually draw. It’s quicker to render in your browser and smaller file size to blur it into a PNG, which only needs to know which colour to draw each pixel, not the instructions for drawing the whole thing from scratch as needed by an SVG.]

The trend is even more obvious when we aggregate up the data to show mean spend for each value of “locations visited”:

This chart also reveals the interesting fact that visitors who report (and remember, it’s only a survey) *zero* locations visited overnight actually have slightly more than average spend; it’s only when we get to the “one location reported” individuals that the pattern becomes strong.

So, the very people who stayed at many locations and hence contribute to the mean spend, length of stay and activities of each location’s visitors, are of course the visitors who spend more money, time and activity while they are in New Zealand. The apparent conundrum of all locations having “better than average” visitors disappears. Nothing to see here, just a warning to look carefully into claims that any particular group is above average, when individuals get to contribute to multiple groups.

Here’s the code for the final two charts:

vw\_IVSSurveyMainHeader %>%

filter(Year >= 2014) %>%

left\_join(places\_visited, by = "SurveyResponseID") %>%

mutate(number\_places\_overnighted = ifelse(is.na(number\_places\_visited), 0, number\_places\_visited)) %>%

ggplot(aes(x = number\_places\_overnighted, y = WeightedSpend, size = PopulationWeight)) +

geom\_point(alpha = 0.1) +

geom\_hline(yintercept = overall\_mean$mean\_spend, colour = "steelblue") +

geom\_smooth(colour = "orange") +

annotate("text", x = 70, y = overall\_mean$mean\_spend - 1000, label = "Overall mean", colour = "steelblue") +

scale\_y\_sqrt("Spend in New Zealand\n(square root transformed scale)",

label = dollar, breaks = c(1000, 10000, 25000, 50000, 100000, 200000)) +

labs(x = "Number of places overnighted in\n",

caption = "Source: MBIE International Visitor Survey") +

theme(legend.position = "none") +

ggtitle("Individuals' spend in New Zealand compared to places visited",

"Data from 2014 to 2017")

vw\_IVSSurveyMainHeader %>%

filter(Year >= 2014) %>%

left\_join(places\_visited, by = "SurveyResponseID") %>%

mutate(number\_places\_overnighted = ifelse(is.na(number\_places\_visited), 0, number\_places\_visited)) %>%

group\_by(number\_places\_overnighted) %>%

summarise(mean\_spend = sum(WeightedSpend \* PopulationWeight) / sum(PopulationWeight),

visitors = sum(PopulationWeight)) %>%

arrange(number\_places\_overnighted) %>%

ggplot(aes(x = number\_places\_overnighted, y = mean\_spend, size = visitors)) +

geom\_hline(yintercept = overall\_mean$mean\_spend, colour = "steelblue") +

geom\_point() +

annotate("text", x = 70, y = overall\_mean$mean\_spend - 500, label = "Overall mean", colour = "steelblue") +

scale\_y\_continuous("Mean spend in New Zealand",

label = dollar) +

labs(x = "Number of places overnighted in\n",

caption = "Source: MBIE International Visitor Survey") +

theme(legend.position = "none") +

ggtitle("Spend in New Zealand compared to places visited",

"Data from 2014 to 2017")

More than ever, my [usual disclaimer](http://ellisp.github.io/about/) applies here – this analysis is my own, not related to employers past and present. If we’d done this analysis at work back when I was with the IVS team, it would have much more peer review and quality control than my humble blog does. If you spot anything, let me know in the comments section.