A system for automated collection of search autocomplete responses and an investigation into the relationship between COVID-19 infection rates and prevalence of COVID-19 related key terms in autocomplete results.

Abstract

This paper is a system to collect Google autocomplete results, demonstrated using an analysis of COVID-19 infection rates and COVID-19 related key terms found in autocomplete results with the stimulus of “test”. Screen scraping was used using the innovative MAS system over 42 different locations worldwide. An VMWare EXSi cluster was to provide computation power for the screen scraping instances while a commercial VPN provider was used to spoof the location (No conflict of interest; service was paid in full). The paper found no correlation (Pearson’s R 0.037) between the prevalence of COVID-19 related key terms and COVID-19 infection rates. This is a statistically significant result with P value below 0.01. This system’s MAS can be used to analyse the filter bubble effect in a future study.

Literature Review

Search engines and their connections are the largest banks of data that humans have ever made. They allow us to find millions of results within seconds, returning links, captions, images, videos, descriptions and more from publicly-indexed websites across the web through many features. One feature that is of particular interest is Autocomplete, also known as word completion or predictive text. It is a feature of major search engines including Google, DuckDuckGo, Bing and Yahoo that predicts the next word to enter into the search box to build a phrase that can then be used to search the database of the search engine to find matching results. Autocomplete is powered by Natural Language Processing (NLP) algorithms that infer intent and meaning from already entered keywords to predict (but not suggest) commonly used keywords that continue to complete the meaning that the searcher is intending to convey. Due to their reliance on NLP algorithms and heuristics from searches, the terms given by autocomplete are thought to be highly dependent on previous searches, however that is something that is discussed in the studies of this literature review and not exhaustively proven, just as with many things related to experimenting on in-production machine learning which has issues due to the black-box nature of machine learning models and the restrictive nature of access to an algorithm which disputably is responsible for

Google's competitive advantage, kept strictly confidential to prevent competitors from improving their respective autocomplete solution.

The literature has identified 2 main problems that have surfaced as the major issues in performing a study on Autocomplete. The first is an issue in the collection of big data sets without using Application Programming Interface (API) methods. This is important as to build studies that conclusively answer questions, many data points are required to establish trends and patterns, and studies that use API methods have no research backing to confirm that they produce the same autocomplete results as a user would receive when typing into a search engine directly. "Game of Missugestions: Semantic Analysis of Search-Autocomplete Manipulations" (Wang 2018) has solved this through the combination of a NLP and machine learning system that achieves a 96.23% precision and 95.63% (Wang 2018, p. 1) recall on terms, all without passing to the Google API except for scraping ancillary data about the page returned by the term for analysis. This technological approach was very interesting, although requiring advanced skills in machine learning and computer science to replicate. This combined with the lack of any published source code means that the paper is of little use to researchers wishing to replicate the study. However, the study's innovative methods worked extremely well, producing results on over 100 million search terms which is a scale that would be impossible by using a screen-scraping or API based approach. This inspired further research into a more technological approach, and to explore a wider range of data collection methods.

MAS is this studies approach to an event-driven mass screen-scraping application with an innovative memory-only Linux Live-CD based solution requiring only an original ISO standard disk drive file and 2 GB of ram per screen-scraping instance. Each instance is controlled by a HTTP API and automation software. Then, the screen-scraping instances are dynamically scaled using a VMWare ESXi cluster to allow a high speed of data collection while still preserving independence between the results.

This solves many of the problems found within the literature, such as "What you see depends on where you sit: The effect of geographical location on web-searching for systematic reviews: A case study" (Chris Cooper 2021), a paper which is looking to determine whether the location of researchers around the world could have any impact on systematic reviews. In (Chris Cooper 2021, p. 2) a single computer is used with Cookies cleared between searches on the chrome browser. However, this method does not prevent persistent fingerprinting via

FLoC, Javascript-based fingerprinting and device fingerprinting which while on their own present no risk to identification when combined can be used to identify the user even after cookie and site data is cleared. The innovative MAS approach proposed by this research paper solves these issues by using Firefox, which contains inbuilt mitigations to prevent device & Javascript de-anonymization as well as a complete reinstall of the OS between search results which ensures advertising identification numbers and other identifiers generated by the computer are reset. Additionally, the media access control address (MAC) is randomised and care is taken to expose no hardware identifiers other than MAC through the hypervisor as would be present on a mobile device. In this way, findings received are fully isolated from each other with a high degree of confidence.

The lower sample size of the paper also brings into question its findings. As such, the MAS system was designed to be scalable and allow a much higher sample size. Of course, analysing autocomplete data is much more simple than analysing full search result data which makes this easy to achieve and practical for a high school student.

Research wise, the paper also confirmed that the idea of testing by location would be viable, the results instilling enough confidence to continue with the project. This idea was further explored in another paper, “Location, Location, Location: The Impact of Geolocation on Web Search Personalization” (Chloe Kliman-Silver 2015) which analysed search results by location in the context of personalization and the “Filter Bubble Effect”, where personalization algorithms prevent users from seeing different perspectives by marking information as not important. This is affected by personalization because demographic traits such as race, income, educational attainment and political affiliation can be inferred from a location. The paper tested a variety of terms including political, controversial and local topics. Logically extending this, MAS intends to extend this to real-world events & further controversial topics such as COVID-19.

The paper is more technically sound than the previous paper, providing evidence that their method prevents personalization on many factors. However, they do not provide any evidence that the mobile and desktop version of Google provide the same results. Additionally, their validation study measuring IP address-based personalization on the mobile browser only showed 94% identical results which should be studied further to come to a statistically significant conclusion. The paper also had an important point which was to

ensure that all queries to Google are sent to the same data centre. This was added to MAS using DNS overrides to ensure consistency.

Research Question

An evaluation of the literature has provided excellent methodological advice that I have incorporated into the MAS system and provided interesting ideas that have given me the confidence to come up with a novel inquiry question:
How much does the prevalence of real-life events affect the algorithms that power Google search?

Hypothesis

The prevalence of Covid-19 related key-terms in autocomplete results correlates with a higher rate of COVID-19 infection in a country.

Methodology

A significant portion of the following methodology contains explanation and discussion of the MAS system. When following these explanations, it can be helpful to view the source code to help build a full understanding of the process. As such, the source code has been published under the GNU General Public License Version 3 (GPLv3) at https://github.com/Mac898/MAS for your convenient access.

The MAS (Macauley Automated Science) system is a screen-scraping based approach to the automatic collection of Google (or with small adaptation) autocomplete results to a search term. Searches are performed in virtual machines using python while connected to VPN servers according to an orchestration system (MASserver). The orchestration system hosts a web server in Python using Flask with routes for different parts of the system to connect to. This orchestration system then keeps the state of the collection process and logs results using a MYSQL (known as DB or database alternatively) database provided through an ORM (Object Relational Mapper) called SQLAlchemy. The orchestration system also contains a basic control panel that can be used to create a new table group in MYSQL, add and remove locations, add and remove stimuli, clear the DB and reset the progress counter for the instances.

The orchestrator controls two additional pieces of software, the VMware EXSi interface (MASexsi) and the collection process tracker (MASclient). Every time a new scraping

operation is started, an instance of MASclient is started and told the ID it needs to respond to via system arguments. The MASclient performs state tracking with the help of the MYSQL database, and through a basic system determines the needed actions to get each test process to completion. The MASexsi system interfaces with the EXSi cluster using the pyvmomi library which uses EXSi's HTTP API. It automatically spins up new VMs consecutively ahead of the injection code, configuring the networking, and booting the OS until the SSH server goes live and it becomes controlled by the MASclient. A list of MAC addresses is kept in the database as "available" nodes which are ready for the client to take on. A program called MASinjector is called by the MASclient and uses the MAC address of the client to lookup the IP address and makes an ssh connection, pushes code, instals basic dependencies to get the code running and executes the experimental code on the node (MASexperiment). This code then pulls down VPN locations, current needs in the database and reserves a location, id, place and Stimuli. The experiment then runs according to the code in MASexperiment and the steps in the state machine. Results are collected and sent back to the MYSQL database.

An experiment to test this system was then needed which resulted in a decision to compare COVID-19 Autocomplete results to COVID-19 Infection rates. This requires a large amount of autocomplete results and so a process was designed to extract these from Google Search. The process starts by installing needed applications and configuring the Linux OS. This includes updating the DNS, updating the apt repositories, installing apt packages and installing python packages. Then the program attempts to start the X server and bring up a working graphical environment using the "startx" command. Then the VPN location is identified and OpenVPN is told to connect to a random server in that country from the VPN provider. Next, Firefox is launched and sent immediately to https://google.com (which is location Generic Google). Then the test starts, and pyautogui is used to write the stimulus into the browser box. Then the screen is captured, the image cropped to only include the search box (Approximately 14% to 70% of the screen horizontally and 58% to 99% vertically, measured from the top left corner), processed using openCV2 algorithms including a resize to 4 times the current screen resolution, A grayscale conversion using the BGR2GRAY transformation (B/W Gamma = 0.299*R + 0.587*G + 0.114*B), A Gaussian blur (Gray type with 3 by 3 blurring, sigmaX and sigmaY of 0) and a thresholding operation in binary inverted mode that converts the pixels to binary data that can be used in the OCR process. The image is run through Google's Tesseract OCR engine using pytesseract to find the final text results. Firefox is then closed, and the VPN disconnected. The system then shuts

down and the resources are freed for MASexsi to spin up a new VM with and the process begins again. Scraping based systems have a high time to capture each result which results in low sample sizes. Therefore, MAS is based on a scalable virtual machine environment to ensure that data can be collected in a way that allows the collection of large sample sizes.

In regards to the experimental method I have taken into account a considerable amount of the research that has already looked into methods to control variables when conducting these kinds of searches. Unfortunately, unlike in searches, no API exists for autocomplete and so results must be screen scraped. Additionally, there is no evidence demonstrating that API results are the same as actual results given by the Google search engine, which further demonstrates the need for a screen scraped approach. Another method presented in the literature was using Machine Learning algorithms and Natural Language Processing (NLP), however this is simply too complicated to achieve with the amount of time and the level of skill of the author.

Keeping this in mind, significant thought was put into ensuring that the browser and its environment is not contaminated. Firefox is used because it has privacy ensuring features that are resistant to Javascript-based fingerprinting, device fingerprinting and FLoC (Federated Learning Of Cohorts) tracking technology. The OS is also completely reinstalled between tests to ensure identifiers in the computer hardware are reset and the MAC addresses are virtual random addresses to ensure no hardware identifiers are trackable. This ensures a high standard that tests are completely isolated from each other.

On the data processing side, a list of key terms have been decided as representative of "COVID-19" in the autocomplete results. These terms include:

tested	positive	testing	isolation	clinic	pcr	covid / covld

Results

ANOVA

Source	SumSquares	df	MeanSquares	F	F Critical (0.01)	IS F > F Critical
Treatment	16.87	41.00	0.45	46.31	1.706	TRUE
Error	1.64	168.00	0.01			
Total	18.51	209.00				
Pearsons (R)	0.037					

See Appendix 1,2,3 for the full dataset. There was no significant correlation between the COVID-19 infection rate and the number of key terms (Pearson's Coefficient 0.037). A F-Critical of 1.706 was calculated for a 0.01 (1/100 chance of statistical fluke) P-Value. The F-value calculated of 46.31 was more than the F-Critical which indicates that the dataset produced a valid result with statistical differences more than P 0.01 between each of the 41 locations (degrees of freedom of treatment group).

Discussion

The experiment has produced conclusive results that show with a high degree of certainty ($P<0.01$) that there is no relationship (Pearson's R 0.037) between the number of COVID-19 related key terms and the COVID-19 infection rate. Statistically, there is no significant correlation amongst all the groups. Despite this, when looking at the regions sorted by highest infection rate, some trends do appear, however these appear to be based more upon the level of COVID-19 in the news and in people's minds. For example, the author's country of Australia has had a detailed imbalance of COVID-19. Perth, Australia had near to zero cases at the date of data collection, while Sydney, Australia has a significant amount more COVID-19 and this is represented in the percentage of key terms as compared to collected results (6.06% of completions in Perth vs 87.88% of completions in Sydney). However, this is a relatively isolated comparison. For example, Seattle, USA had a rate of 17.24% but had 1196 more cases per day than Sydney. Or Houston, Texas had 39.39% but had 6416 more cases per day. As such, there is clearly another variable involved that was not controlled for in this dataset.

Despite the result having high certainty, there are still the possibilities of systematic error. One possibility for this systematic error is that of uncontrolled variables. As discussed in the literature review, best attempts have been made to account for all variables that would otherwise differ the groups. It is therefore the opinion of this paper that either a multivariate relationship applies where, for example, prevalence in conventional and online media, and public conscience or there is a relationship with a different independent variable. It may be considered that these are affected by the infection rate which is true, and evidenced by the pearson's coefficient's slight positive value of 0.037. However, there are simply too many intermediate variables between the infection rate and prevalence of autocomplete results.

However, it is clear that the MAS software worked as planned with the high number of samples (Appendix 3: 934 Samples) collected over 30 minutes, with a P value well below 0.01. This indicates that the programmatic portion of this paper was a statistically significant success, and could be used to conduct other experiments. This is a significant way to run an experiment of this kind and further confirms existing research such as “Location, Location, Location: The Impact of Geolocation on Web Search Personalization” (Chloe Kliman-Silver 2015).

The author has also put significant effort into ensuring that the experimental system can be easily modified for a variety of experiments. With some basic Linux experience, almost any screen scraping or other task can be accommodated in the code. To modify the code, MASexperiment should be rewritten to accomplish your goal within the virtual machine. Then MASclient should hook your stages into the MASexperiment code. Then configure settings.json with your required packages to start the python program, the stages and their execution order, the EXSi settings and the experimental details and add the VPN credentials to the appropriate section. Then, modify the regions to an appropriate format for your VPN provider (See the MASexperiment code that gets passed this name chosen per instance. Then modify the MASinjector to upload your code, which you should have zipped as a .tar file. The code is run using Python 3.5 however this can also be changed inside the MASinjector. Following these steps will allow you to run your own experiment with the MAS system.

With this success in mind, possible further research for the COVID-19 hypothesis would be to look more at the search results themselves to see if patterns or trends can be found in these results, testing the variable of online media for example. However, this would require a more complex analysis method which a more capable researcher may be able to contribute. Additionally, this could be used to perform an analysis of the filter bubble effect. With the high level of scalability, this could be a deep look into multiple different categories of searches such as politics, food venues, entertainment, science and social media.

Conclusion

In conclusion, the paper has successfully demonstrated at a high confidence level that there is no significant correlation between the infection rate and percentage of key terms in autocomplete results (Pearson’s R 0.037). This was a statistically significant result with P

value less than 0.01. Variables were well controlled in the collection process using a complex but comprehensive methodology that ensures fair data collection. However, multiple variables influence the prevalence of autocomplete results, and the infection rate variable does not well correlate to this result. Other variables including the amount of coverage on conventional and internet media, the public consciousness etc may have a large effect on the result more directly, however even this is speculation. As mentioned in the literature review, the Google autocomplete algorithm is a trade secret and studying it for its correlations is difficult. Further research is needed into these variables to determine possible correlations. The MAS system can be used to help with this further research with simple modifications as discussed in the methodology.

Reference List

Danny S 2018, 'How Google autocomplete works in Search', Google: The Keyword, accessed 22 Feb 2022, <https://blog.google/products/search/how-google-autocomplete-works-search/>

Chloe K, Aniko H, David L, Christo W 2015, 'Location, Location, Location: The Impact of Geolocation on Web Search Personalization', *ResearchGate,* accessed 22 Feb 2022, <https://www.researchgate.net/publication/301417602_Location_Location_Location>

Chris C, Theo L, Ute S 2021, 'What you see depends on where you sit: The effect of geographical location on web-searching for systematic reviews: A case study', Research Synthesis Methods, Volume 12 Issue 4, P. 557-570, accessed 22 Feb 2022, <https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1485>

Peng W, Xianghang M, Xiaojing L, XiaoFeng W, Kan Y, Feng Q, Raheem B 2018, 'Game of Missuggestions: Semantic Analysis of Search-Autocomplete Manipulations', accessed 22 Feb 2022, <https://homes.luddy.indiana.edu/xw7/papers/peng18ndss.pdf>

Appendices

Appendix 1 - Data

ID	Count	Samples	Percentage	Location ID	Stimulus	Location	SSER
69	4	7	0.5714285714	1	test	AL-Tirana-tia	0.005102040816
84	4	7	0.5714285714	1	test	AL-Tirana-tia	0.005102040816
92	6	8	0.75	1	test	AL-Tirana-tia	0.01147959184
160	4	7	0.5714285714	1	test	AL-Tirana-tia	0.005102040816
168	6	8	0.75	1	test	AL-Tirana-tia	0.01147959184
98	4	7	0.5714285714	2	test	AR-Buenos-Aires-eze	0
119	4	7	0.5714285714	2	test	AR-Buenos-Aires-eze	0
121	4	7	0.5714285714	2	test	AR-Buenos-Aires-eze	0
181	4	7	0.5714285714	2	test	AR-Buenos-Aires-eze	0
203	4	7	0.5714285714	2	test	AR-Buenos-Aires-eze	0
11	7	8	0.875	3	test	AU-Adelaide-adl	0
31	7	8	0.875	3	test	AU-Adelaide-adl	0
66	7	8	0.875	3	test	AU-Adelaide-adl	0
188	7	8	0.875	3	test	AU-Adelaide-adl	0
205	7	8	0.875	3	test	AU-Adelaide-adl	0
59	7	8	0.875	4	test	AU-Brisbane-bne	0
64	7	8	0.875	4	test	AU-Brisbane-bne	0
94	7	8	0.875	4	test	AU-Brisbane-bne	0
127	7	8	0.875	4	test	AU-Brisbane-bne	0
179	7	8	0.875	4	test	AU-Brisbane-bne	0
53	5	8	0.625	5	test	AU-Melbourne-mel	0.09
78	1	8	0.125	5	test	AU-Melbourne-mel	0.04
102	5	8	0.625	5	test	AU-Melbourne-mel	0.09
132	1	8	0.125	5	test	AU-Melbourne-mel	0.04
151	1	8	0.125	5	test	AU-Melbourne-mel	0.04
2	2	8	0.25	6	test	AU-Perth-per	0.04
28	0	8	0	6	test	AU-Perth-per	0.0025
83	0	8	0	6	test	AU-Perth-per	0.0025
183	0	8	0	6	test	AU-Perth-per	0.0025
195	0	8	0	6	test	AU-Perth-per	0.0025
5	7	8	0.875	7	test	AU-Sydney-syd	0.005625
38	6	8	0.75	7	test	AU-Sydney-syd	0.0025
71	6	8	0.75	7	test	AU-Sydney-syd	0.0025
128	7	8	0.875	7	test	AU-Sydney-syd	0.005625
149	6	8	0.75	7	test	AU-Sydney-syd	0.0025
50	0	6	0	8	test	BR-Sao-Paulo-gru	0.06612244898
55	3	7	0.4285714286	8	test	BR-Sao-Paulo-gru	0.0293877551
110	3	7	0.4285714286	8	test	BR-Sao-Paulo-gru	0.0293877551
159	3	7	0.4285714286	8	test	BR-Sao-Paulo-gru	0.0293877551
163	0	6	0	8	test	BR-Sao-Paulo-gru	0.06612244898
82	4	7	0.5714285714	9	test	CA-Montreal-yul	0
109	4	7	0.5714285714	9	test	CA-Montreal-yul	0
118	4	7	0.5714285714	9	test	CA-Montreal-yul	0
138	4	7	0.5714285714	9	test	CA-Montreal-yul	0
169	4	7	0.5714285714	9	test	CA-Montreal-yul	0
42	3	8	0.375	10	test	CA-Toronto-tor	0.001225
47	3	8	0.375	10	test	CA-Toronto-tor	0.001225
91	1	5	0.2	10	test	CA-Toronto-tor	0.0196
175	3	8	0.375	10	test	CA-Toronto-tor	0.001225

191	3	8	0.375	10	test	CA-Toronto-tor	0.001225
43	0	8	0	11	test	CA-Vancouver-yvr	0
101	0	8	0	11	test	CA-Vancouver-yvr	0
103	0	8	0	11	test	CA-Vancouver-yvr	0
137	0	8	0	11	test	CA-Vancouver-yvr	0
150	0	8	0	11	test	CA-Vancouver-yvr	0
7	2	7	0.2857142857	12	test	CL-Santiago-scl	0
12	2	7	0.2857142857	12	test	CL-Santiago-scl	0
16	2	7	0.2857142857	12	test	CL-Santiago-scl	0
193	2	7	0.2857142857	12	test	CL-Santiago-scl	0
200	2	7	0.2857142857	12	test	CL-Santiago-scl	0
21	1	7	0.1428571429	13	test	CR-San-Jose-sjo	0.00005102040816
107	1	7	0.1428571429	13	test	CR-San-Jose-sjo	0.00005102040816
125	0	7	0	13	test	CR-San-Jose-sjo	0.01841836735
190	2	8	0.25	13	test	CR-San-Jose-sjo	0.01306122449
202	1	7	0.1428571429	13	test	CR-San-Jose-sjo	0.00005102040816
8	0	8	0	14	test	IN-Mumbai-bom	0
14	0	7	0	14	test	IN-Mumbai-bom	0
57	0	7	0	14	test	IN-Mumbai-bom	0
167	0	7	0	14	test	IN-Mumbai-bom	0
180	0	7	0	14	test	IN-Mumbai-bom	0
15	0	8	0	15	test	IN-New-Delhi-del	0
30	0	8	0	15	test	IN-New-Delhi-del	0
77	0	8	0	15	test	IN-New-Delhi-del	0
147	0	8	0	15	test	IN-New-Delhi-del	0
192	0	8	0	15	test	IN-New-Delhi-del	0
1	0	7	0	16	test	JP-Tokyo-nrt	0
120	0	7	0	16	test	JP-Tokyo-nrt	0
124	0	6	0	16	test	JP-Tokyo-nrt	0
133	0	7	0	16	test	JP-Tokyo-nrt	0
161	0	7	0	16	test	JP-Tokyo-nrt	0
45	0	6	0	17	test	KR-Seoul-sel	0
68	0	6	0	17	test	KR-Seoul-sel	0
80	0	6	0	17	test	KR-Seoul-sel	0
141	0	6	0	17	test	KR-Seoul-sel	0
173	0	6	0	17	test	KR-Seoul-sel	0
70	6	7	0.8571428571	18	test	MD-Chisinau-kiv	0
88	6	7	0.8571428571	18	test	MD-Chisinau-kiv	0
97	6	7	0.8571428571	18	test	MD-Chisinau-kiv	0
143	6	7	0.8571428571	18	test	MD-Chisinau-kiv	0
164	6	7	0.8571428571	18	test	MD-Chisinau-kiv	0
89	0	8	0	19	test	MX-Guadalajara-gdl	0
90	0	8	0	19	test	MX-Guadalajara-gdl	0
93	0	8	0	19	test	MX-Guadalajara-gdl	0
207	0	8	0	19	test	MX-Guadalajara-gdl	0
208	0	8	0	19	test	MX-Guadalajara-gdl	0
34	2	8	0.25	20	test	MY-Kuala-Lumpur-kul	0
79	2	8	0.25	20	test	MY-Kuala-Lumpur-kul	0
87	2	8	0.25	20	test	MY-Kuala-Lumpur-kul	0
135	2	8	0.25	20	test	MY-Kuala-Lumpur-kul	0

176	2	8	0.25	20	test	MY-Kuala-Lumpur-kul	0
13	5	8	0.625	21	test	NZ-Auckland-akl	0
49	5	8	0.625	21	test	NZ-Auckland-akl	0
100	5	8	0.625	21	test	NZ-Auckland-akl	0
145	5	8	0.625	21	test	NZ-Auckland-akl	0
162	5	8	0.625	21	test	NZ-Auckland-akl	0
60	2	7	0.2857142857	22	test	PE-Lima-lim	0
67	2	7	0.2857142857	22	test	PE-Lima-lim	0
96	2	7	0.2857142857	22	test	PE-Lima-lim	0
140	2	7	0.2857142857	22	test	PE-Lima-lim	0
170	2	7	0.2857142857	22	test	PE-Lima-lim	0
17	4	7	0.5714285714	23	test	RS-Belgrade-beg	0
19	4	7	0.5714285714	23	test	RS-Belgrade-beg	0
22	4	7	0.5714285714	23	test	RS-Belgrade-beg	0
134	4	7	0.5714285714	23	test	RS-Belgrade-beg	0
157	4	7	0.5714285714	23	test	RS-Belgrade-beg	0
4	1	8	0.125	24	test	SG-Singapore-sin	0.09
23	4	8	0.5	24	test	SG-Singapore-sin	0.005625
105	4	8	0.5	24	test	SG-Singapore-sin	0.005625
129	4	8	0.5	24	test	SG-Singapore-sin	0.005625
148	4	8	0.5	24	test	SG-Singapore-sin	0.005625
35	0	7	0	25	test	SI-Ljubljana-lju	0
81	0	7	0	25	test	SI-Ljubljana-lju	0
114	0	7	0	25	test	SI-Ljubljana-lju	0
146	0	7	0	25	test	SI-Ljubljana-lju	0
186	0	7	0	25	test	SI-Ljubljana-lju	0
56	1	7	0.1428571429	26	test	US-Ashburn-iad	0
63	1	7	0.1428571429	26	test	US-Ashburn-iad	0
95	1	7	0.1428571429	26	test	US-Ashburn-iad	0
194	1	7	0.1428571429	26	test	US-Ashburn-iad	0
198	1	7	0.1428571429	26	test	US-Ashburn-iad	0
9	0	7	0	27	test	US-Atlanta-atl	0
20	0	7	0	27	test	US-Atlanta-atl	0
37	0	7	0	27	test	US-Atlanta-atl	0
171	0	7	0	27	test	US-Atlanta-atl	0
196	0	7	0	27	test	US-Atlanta-atl	0
6	4	7	0.5714285714	28	test	US-Boston-bos	0.0001147959184
27	4	7	0.5714285714	28	test	US-Boston-bos	0.0001147959184
36	5	8	0.625	28	test	US-Boston-bos	0.001836734694
142	4	7	0.5714285714	28	test	US-Boston-bos	0.0001147959184
189	4	7	0.5714285714	28	test	US-Boston-bos	0.0001147959184
3	2	8	0.25	29	test	US-Charlotte-clt	0.0025
86	2	8	0.25	29	test	US-Charlotte-clt	0.0025
99	4	8	0.5	29	test	US-Charlotte-clt	0.04
152	0	8	0	29	test	US-Charlotte-clt	0.09
177	4	8	0.5	29	test	US-Charlotte-clt	0.04
18	3	8	0.375	30	test	US-Chicago-chi	0
54	3	8	0.375	30	test	US-Chicago-chi	0
72	3	8	0.375	30	test	US-Chicago-chi	0
139	3	8	0.375	30	test	US-Chicago-chi	0

165	3	8	0.375	30	test	US-Chicago-chi	0
32	0	8	0	31	test	US-Cincinnati-cvg	0
76	0	8	0	31	test	US-Cincinnati-cvg	0
122	0	8	0	31	test	US-Cincinnati-cvg	0
197	0	8	0	31	test	US-Cincinnati-cvg	0
199	0	8	0	31	test	US-Cincinnati-cvg	0
40	0	8	0	32	test	US-Dallas-dal	0
41	0	8	0	32	test	US-Dallas-dal	0
46	0	8	0	32	test	US-Dallas-dal	0
144	0	8	0	32	test	US-Dallas-dal	0
184	0	8	0	32	test	US-Dallas-dal	0
26	3	8	0.375	33	test	US-Denver-den	0
51	3	8	0.375	33	test	US-Denver-den	0
112	3	8	0.375	33	test	US-Denver-den	0
130	3	8	0.375	33	test	US-Denver-den	0
156	3	8	0.375	33	test	US-Denver-den	0
33	3	8	0.375	34	test	US-Houston-hou	0
52	3	8	0.375	34	test	US-Houston-hou	0
74	3	8	0.375	34	test	US-Houston-hou	0
172	3	8	0.375	34	test	US-Houston-hou	0
204	3	8	0.375	34	test	US-Houston-hou	0
39	6	8	0.75	35	test	US-Las-Vegas-las	0.09
62	2	8	0.25	35	test	US-Las-Vegas-las	0.04
65	2	8	0.25	35	test	US-Las-Vegas-las	0.04
158	6	8	0.75	35	test	US-Las-Vegas-las	0.09
182	2	8	0.25	35	test	US-Las-Vegas-las	0.04
61	6	8	0.75	36	test	US-Los-Angeles-lax	0
75	6	8	0.75	36	test	US-Los-Angeles-lax	0
106	6	8	0.75	36	test	US-Los-Angeles-lax	0
131	6	8	0.75	36	test	US-Los-Angeles-lax	0
166	6	8	0.75	36	test	US-Los-Angeles-lax	0
104	5	7	0.7142857143	37	test	US-Miami-mia	0.001275510204
115	5	7	0.7142857143	37	test	US-Miami-mia	0.001275510204
126	5	8	0.625	37	test	US-Miami-mia	0.002869897959
201	5	8	0.625	37	test	US-Miami-mia	0.002869897959
206	5	7	0.7142857143	37	test	US-Miami-mia	0.001275510204
10	2	8	0.25	38	test	US-New-Orleans-msy	0
111	0	10	0	38	test	US-New-Orleans-msy	0.0625
113	2	8	0.25	38	test	US-New-Orleans-msy	0
155	3	8	0.375	38	test	US-New-Orleans-msy	0.015625
174	3	8	0.375	38	test	US-New-Orleans-msy	0.015625
24	4	7	0.5714285714	39	test	US-New-York-nyc	0.01306122449
85	4	7	0.5714285714	39	test	US-New-York-nyc	0.01306122449
123	4	7	0.5714285714	39	test	US-New-York-nyc	0.01306122449
209	4	7	0.5714285714	39	test	US-New-York-nyc	0.01306122449
210	0	3	0	39	test	US-New-York-nyc	0.2089795918
25	0	7	0	40	test	US-Phoenix-phx	0
29	0	7	0	40	test	US-Phoenix-phx	0
44	0	7	0	40	test	US-Phoenix-phx	0
178	0	7	0	40	test	US-Phoenix-phx	0

187	0	7	0	40	test	US-Phoenix-phx	0
108	0	8	0	41	test	US-San-Jose-sjc	0
116	0	8	0	41	test	US-San-Jose-sjc	0
117	0	8	0	41	test	US-San-Jose-sjc	0
153	0	8	0	41	test	US-San-Jose-sjc	0
154	0	8	0	41	test	US-San-Jose-sjc	0
48	1	7	0.1428571429	42	test	US-Seattle-sea	0
58	1	7	0.1428571429	42	test	US-Seattle-sea	0
73	1	7	0.1428571429	42	test	US-Seattle-sea	0
136	1	7	0.1428571429	42	test	US-Seattle-sea	0
185	1	7	0.1428571429	42	test	US-Seattle-sea	0

Appendix 2 - Summarised Data

Location ID	Locations	Mean	Sample Size	SSR	SSE	SST
1	AL-Tirana-tia	0.643	5	0.507	0.03826530612	0.546
2	AR-Buenos-Aires-eze	0.571	5	0.305	0	0.305
3	AU-Adelaide-adl	0.875	5	1.516	0	1.516
4	AU-Brisbane-bne	0.875	5	1.516	0	1.516
5	AU-Melbourne-mel	0.325	5	0.000	0.3	0.300
6	AU-Perth-per	0.050	5	0.376	0.05	0.426
7	AU-Sydney-syd	0.800	5	1.131	0.01875	1.150
8	BR-Sao-Paulo-gru	0.257	5	0.023	0.2204081633	0.243
9	CA-Montreal-yul	0.571	5	0.305	0	0.305
10	CA-Toronto-tor	0.340	5	0.001	0.0245	0.026
11	CA-Vancouver-yvr	0.000	5	0.526	0	0.526
12	CL-Santiago-scl	0.286	5	0.007	0	0.007
13	CR-San-Jose-sjo	0.136	5	0.178	0.03163265306	0.210
14	IN-Mumbai-bom	0.000	5	0.526	0	0.526
15	IN-New-Delhi-del	0.000	5	0.526	0	0.526
16	JP-Tokyo-nrt	0.000	5	0.526	0	0.526
17	KR-Seoul-sel	0.000	5	0.526	0	0.526
18	MD-Chisinau-kiv	0.857	5	1.419	0	1.419
19	MX-Guadalajara-gdl	0.000	5	0.526	0	0.526
20	MY-Kuala-Lumpur-kul	0.250	5	0.028	0	0.028
21	NZ-Auckland-akl	0.625	5	0.452	0	0.452
22	PE-Lima-lim	0.286	5	0.007	0	0.007
23	RS-Belgrade-beg	0.571	5	0.305	0	0.305
24	SG-Singapore-sin	0.425	5	0.051	0.1125	0.163
25	SI-Ljubljana-lju	0.000	5	0.526	0	0.526
26	US-Ashburn-iad	0.143	5	0.165	0	0.165
27	US-Atlanta-atl	0.000	5	0.526	0	0.526
28	US-Boston-bos	0.582	5	0.332	0.002295918367	0.335
29	US-Charlotte-clt	0.300	5	0.003	0.175	0.178
30	US-Chicago-chi	0.375	5	0.013	0	0.013
31	US-Cincinnati-cvg	0.000	5	0.526	0	0.526
32	US-Dallas-dal	0.000	5	0.526	0	0.526
33	US-Denver-den	0.375	5	0.013	0	0.013
34	US-Houston-hou	0.375	5	0.013	0	0.013
35	US-Las-Vegas-las	0.450	5	0.079	0.3	0.379
36	US-Los-Angeles-lax	0.750	5	0.906	0	0.906
37	US-Miami-mia	0.679	5	0.627	0.009566326531	0.637
38	US-New-Orleans-msy	0.250	5	0.028	0.09375	0.121
39	US-New-York-nyc	0.457	5	0.088	0.2612244898	0.349
40	US-Phoenix-phx	0.000	5	0.526	0	0.526
41	US-San-Jose-sjc	0.000	5	0.526	0	0.526
42	US-Seattle-sea	0.143	5	0.165	0	0.165
	Overall	0.324	210	16.872	1.637892857	18.510

Appendix 3 - Pearson's Correlation Dataset

Location	Count	Samples	Percentage	InfectionRate
US-Seattle-sea	5	29	17.24%	14785.80706
US-New-York-nyc	13	25	52.00%	23956.74849
US-New-Orleans-msy	9	35	25.71%	20314.29421
US-Miami-mia	22	31	70.97%	41532.3857
US-Los-Angeles-lax	26	33	78.79%	26002.16334
US-Las-Vegas-las	14	33	42.42%	21542.14359
US-Houston-hou	13	33	39.39%	20005.66479
US-Denver-den	13	33	39.39%	21639.24913
US-Chicago-chi	13	33	39.39%	20879.56021
US-Charlotte-clt	12	33	36.36%	23119.07172
US-Boston-bos	18	30	60.00%	24508.43195
US-Ashburn-iad	5	29	17.24%	14912.05161
SG-Singapore-sin	17	33	51.52%	5788.122167
RS-Belgrade-beg	17	29	58.62%	13700.71314
PE-Lima-lim	9	29	31.03%	12191.965
NZ-Auckland-akl	21	33	63.64%	334.8241365
MY-Kuala-Lumpur-kul	10	33	30.30%	12324.67386
MD-Chisinau-kiv	27	29	93.10%	10772.26043
CR-San-Jose-sjo	5	30	16.67%	13397.42299
CL-Santiago-scl	9	29	31.03%	12100.88899
CA-Toronto-tor	11	30	36.67%	6942.958206
CA-Montreal-yul	17	29	58.62%	9992.405426
BR-Sao-Paulo-gru	7	28	25.00%	10069.93198
AU-Sydney-syd	29	33	87.88%	13589.78813
AU-Perth-per	2	33	6.06%	55.57667452
AU-Melbourne-mel	10	33	30.30%	12528.15276
AU-Brisbane-bne	33	33	100.00%	7625.960317
AU-Adelaide-adl	33	33	100.00%	6564.076288
AR-Buenos-Aires-eze	17	29	58.62%	18301.78957
AL-Tirana-tia	24	31	77.42%	8830.565015
Totals	461	934		