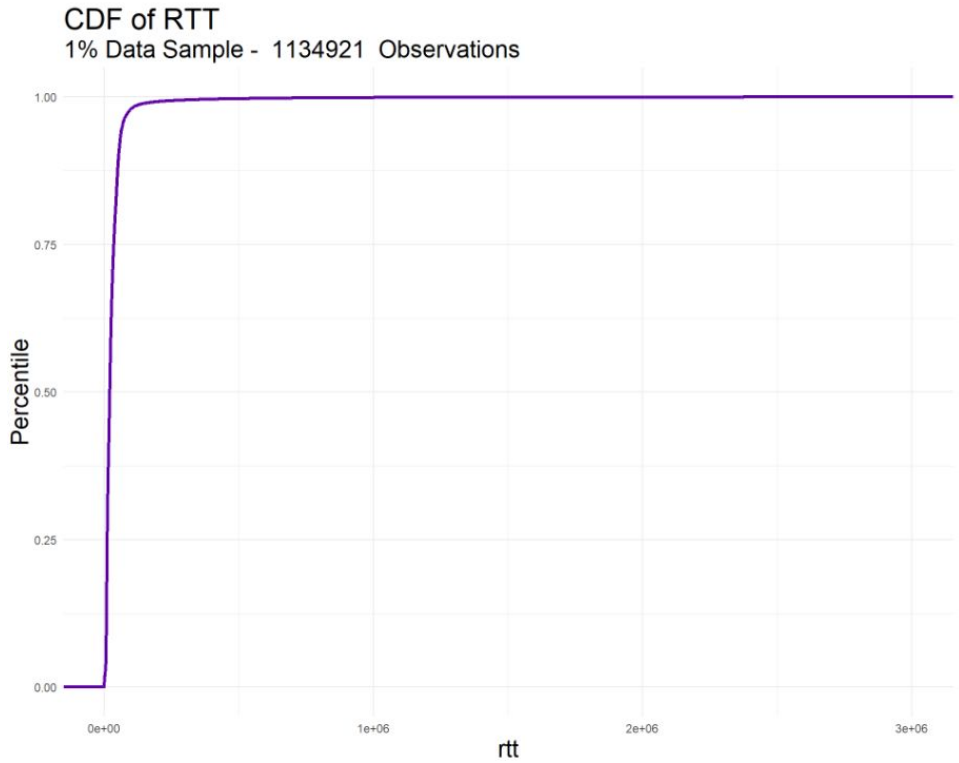
COMP 590-042
Project 1

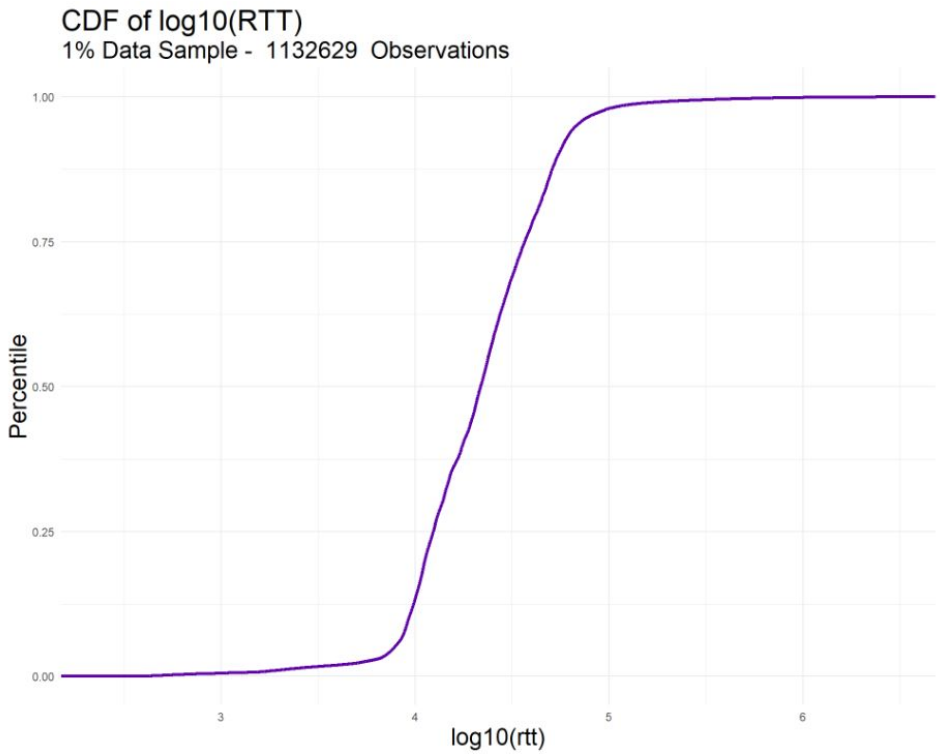
Group Members:
Benjamin Lu, Joshua Shrestha
Angello Polo Espinoza, Jeremy Chao

Introduction

The FCC distributes black boxes that periodically issue DNS queries and record the latency of the responses. In our project, we will attempt to extract meaningful conclusions from differences in DNS resolution time (rtt) across different geographic locations (regions and states), times of day, and internet service providers (ISPs) and connection types. We will use the MapReduce programming framework with the Hadoop file system in order to extract the desired data and plot our results in R with the ggplot2 package. We will sample 1% of the total 13 GB dns.csv for each question for plotting purposes.



For context, we provide a CDF plot of rtt's from a 1% sample of the data above. The rtt's are heavily skewed right with large outliers and an interdecile range of approximately ~0 to ~1e+05 microseconds. As a result, the plot is very difficult to read. For the remainder of this report, our CDFs will be reported on a logarithmic scale (as in the plot below) so that we can easily spot the differences between plots and better draw conclusions from them. This scale also allows us to easily determine the rtt at a given percentile with a simple exponent operation.



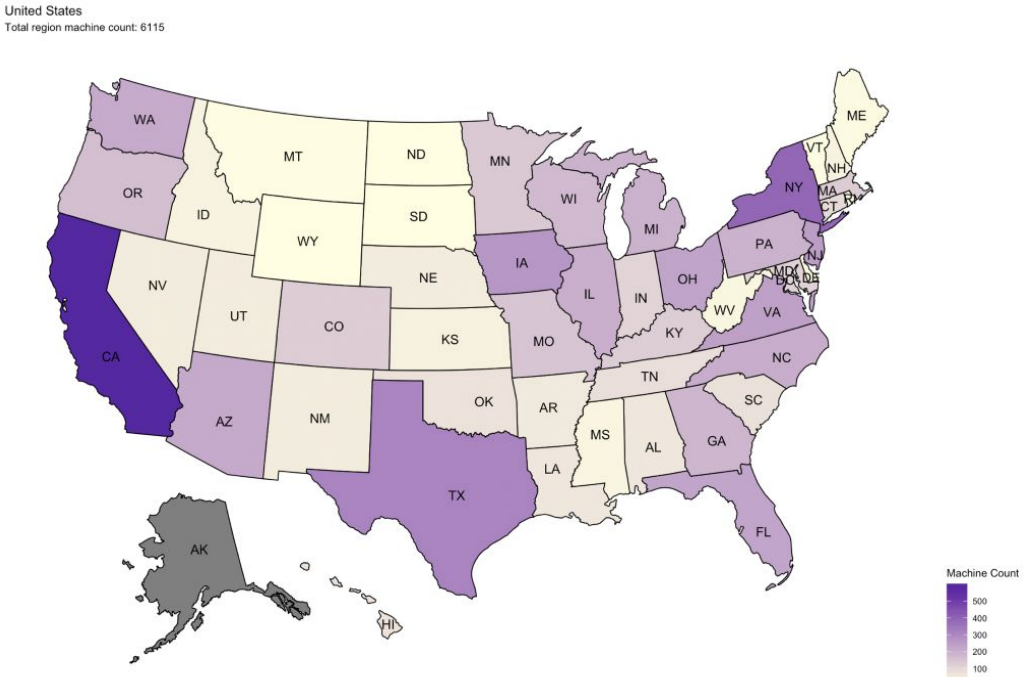
We will perform our analysis in the following manner: exploring how location affects DNS performance, exploring how time affects DNS performance, and exploring how ISP and connection type affect DNS performance. By the end, we will have a comprehensive analysis that dives into each of these points to bring a broader understanding of the factors that do and do not affect DNS performance. We will measure DNS performance using rtt.

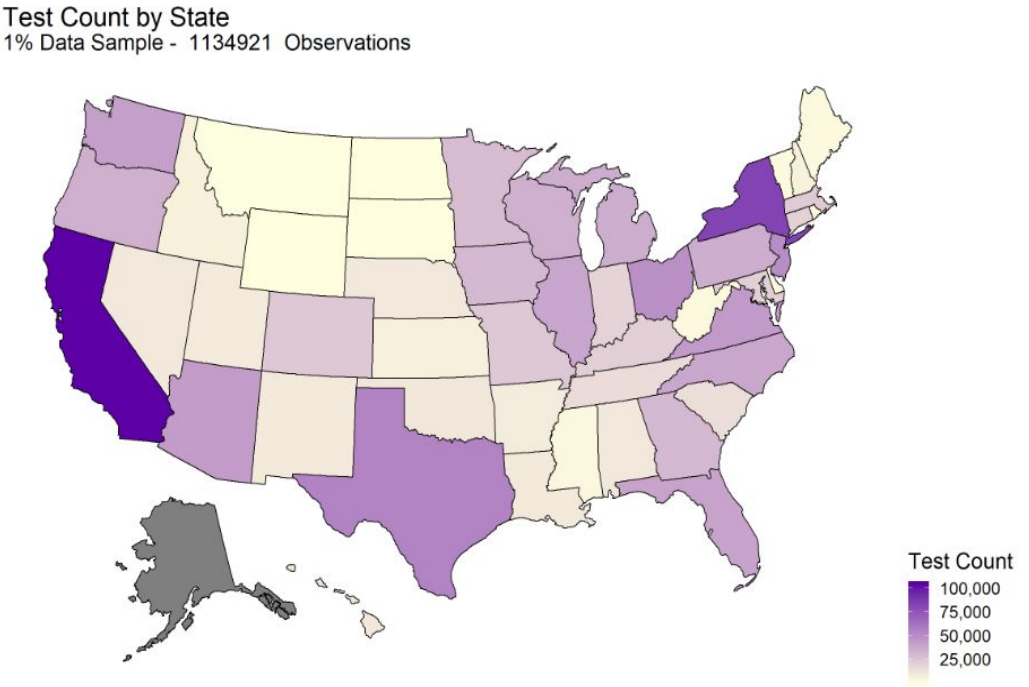
PART I

How does DNS performance differ by location?

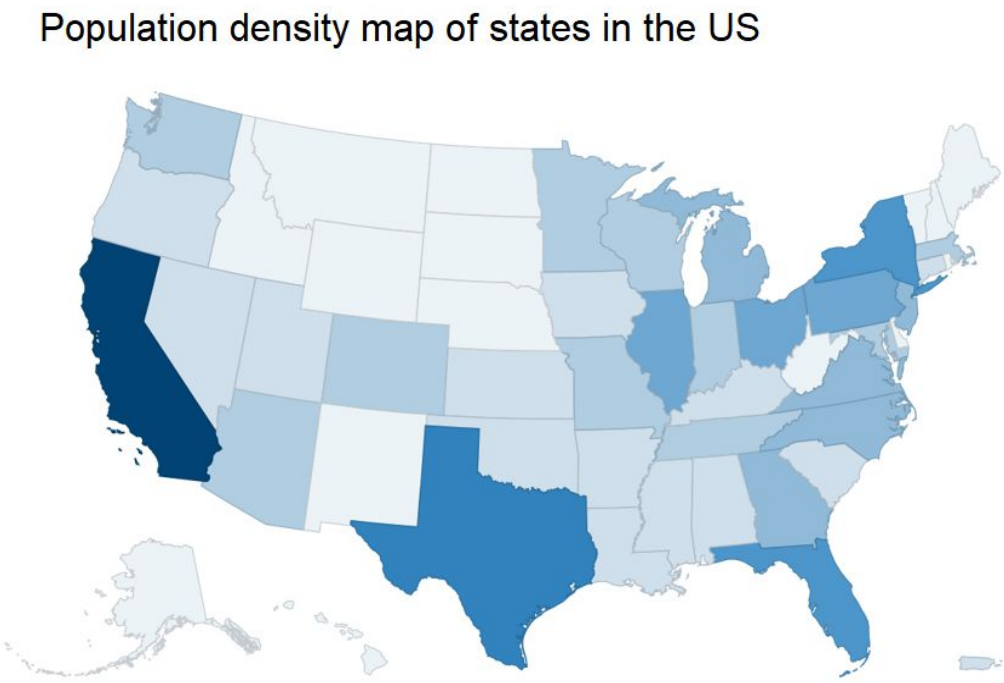
The data indicates the state and the region in which different machines are located. We use this data to guide our analysis of how DNS performance and rtt might differ depending on location. For context, we will first explore how black boxes are distributed across states and draw conclusions about how this might affect test counts per state. We will then explore how rtt differs by region and state.

1. How are the machines distributed around the country?





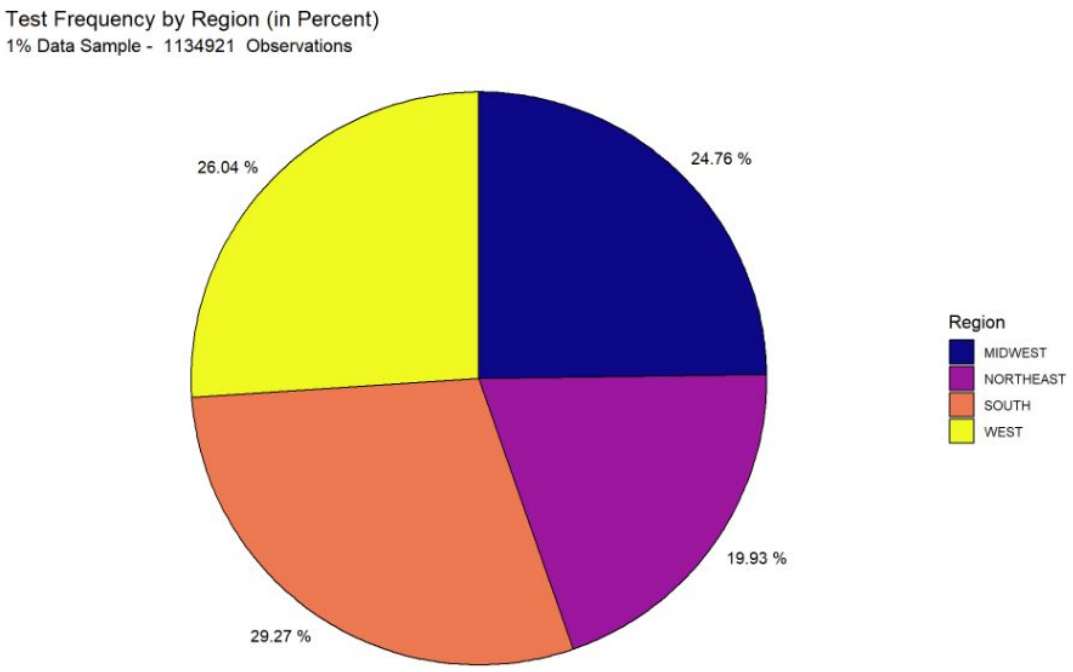
We ask this question to gain a general sense of how test metrics may be skewed towards certain locations or people that live in certain areas. In reviewing the plots, it seems that states with large populations such as California, Texas, Virginia, and New York have high concentrations of black boxes and host more tests. This is confirmed by noting the similarities between the charts with "Population density map of states in the US." This would suggest that these black boxes were distributed randomly to the U.S. population rather distributed evenly across different geographic locations. This suggest that boxes in less populous states do not perform any more tests than boxes in more populous states. While this will provide a good picture of average DNS performance across the population, sparsely populated states will have far less influence on test metrics coming from this data set.



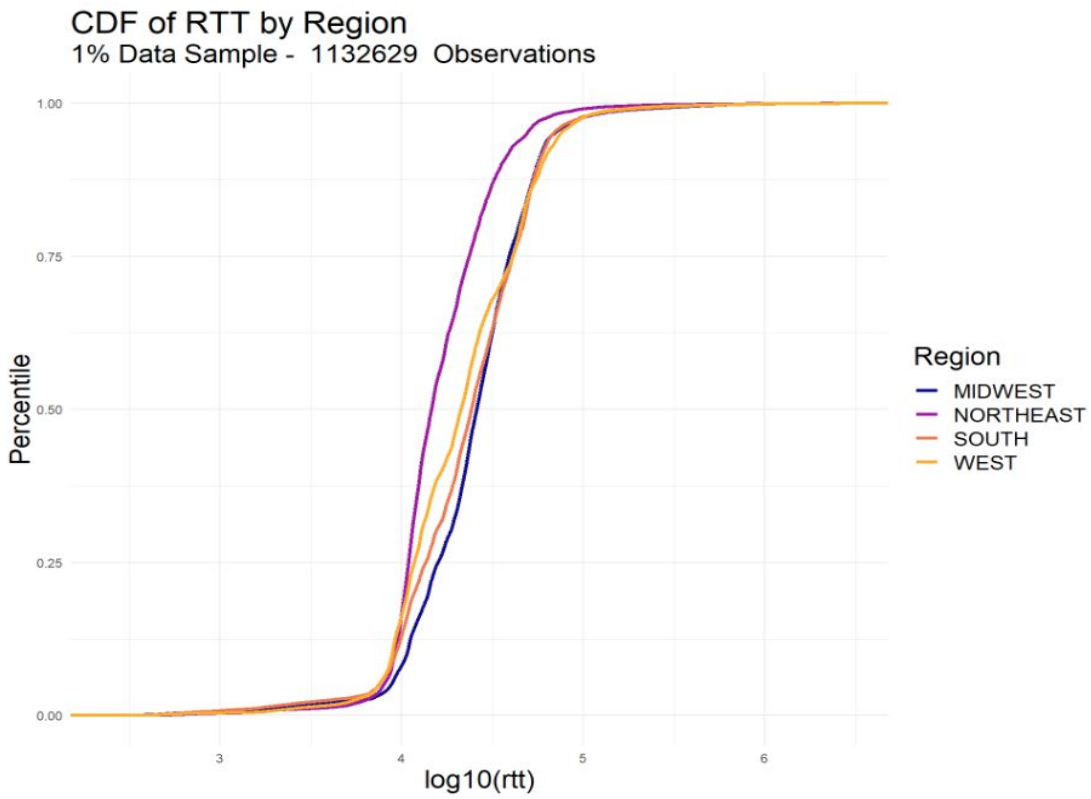
2. How do resolution times differ by location?

To field our approach to this question, we started out with a pie chart on the following page of the test frequency by region sampled on 1% of our data. Here we see that the test frequency is split fairly evenly, with differences seeming to be attributed to the relative population of each region, which is consistent with our findings above. Hence, the South has the highest test frequency because it has the largest population as seen in the table below.

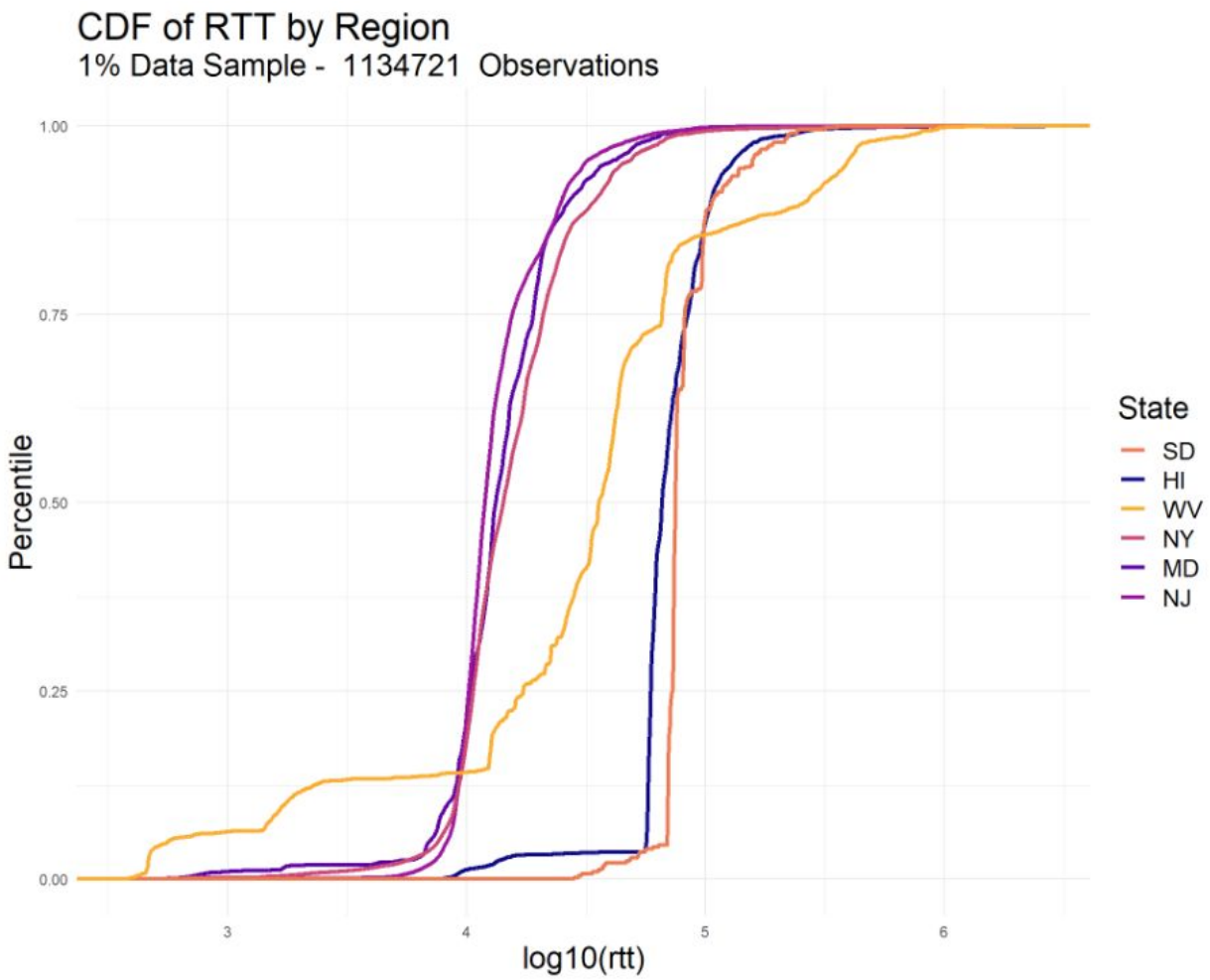
Region	Population	Percentage
Northeast	56,111,079	17.2%
Midwest	68,308,744	20.9%
West	77,993,663	23.8%
South	124,753,948	38.1%



We would now like to test whether there are significant differences in rtt across different regions. We show these rtt's in the below CDF of rtt by region sampled on 1% of our data.



We see that the Northeast, having the lowest test frequency, also has the lowest median rtt. With the Midwest having the second lowest test frequency, it would be reasonable to assume that it would have the second lowest median rtt. However, the Midwest has the highest median rtt. Therefore, regional differences in test frequency, and therefore regional population, does not seem to correlate with differences in median rtt.



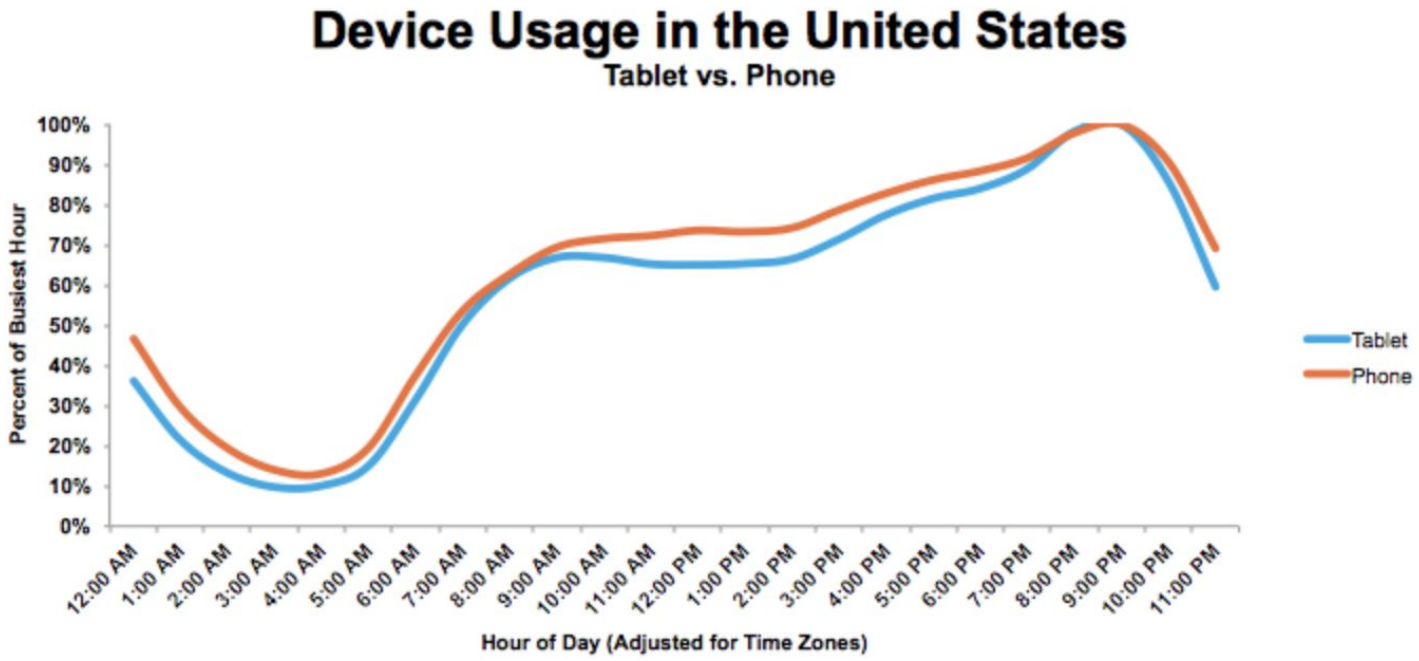
To further investigate, we created a CDF of the rtt's of states with the 3 highest and 3 lowest mean rtt's. The states with the lowest mean rtt were NY, MD, and NJ while the states with the highest were SD, HI, and WV. The 3 lowest states have similar CDF plots with the median around ~10^4.2 microseconds. Interestingly, these states rank in the highest population densities per state in the US (NY 8th, MD 6th, NJ 2nd), indicating the presence of an inverse correlation between population density and rtt. Similarly, the highest 3 states do not rank as high in population density (SD 46th, HI 19th, WV 35th)[1] which could contribute to their higher median rtt's.

We found that regional test frequencies, and therefore regional populations, did not seem to correlate with median rtt. Concluded that regions were far too abstract, we dropped down to states while changing our measure to population density to account for differences in state size. Population density seems to have a significant correlation to rtt's across individual states. Higher population densities tend to be correlated with improved DNS performance.

PART II

How does time affect DNS performance?

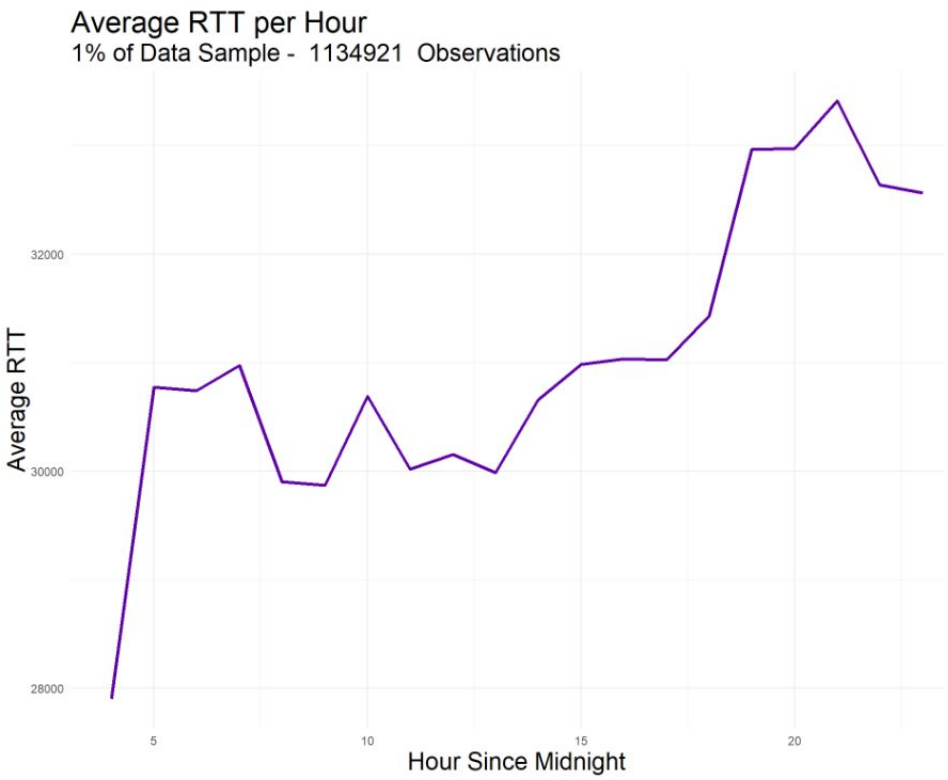
The data has the date and time in UTC of when the test finished for each DNS request. We use this data to explore trends related to the effects time of day and day of week have on DNS performance. For reference, we will refer to the below plot for general American internet usage trends over the course of a day, taken from Localytics.

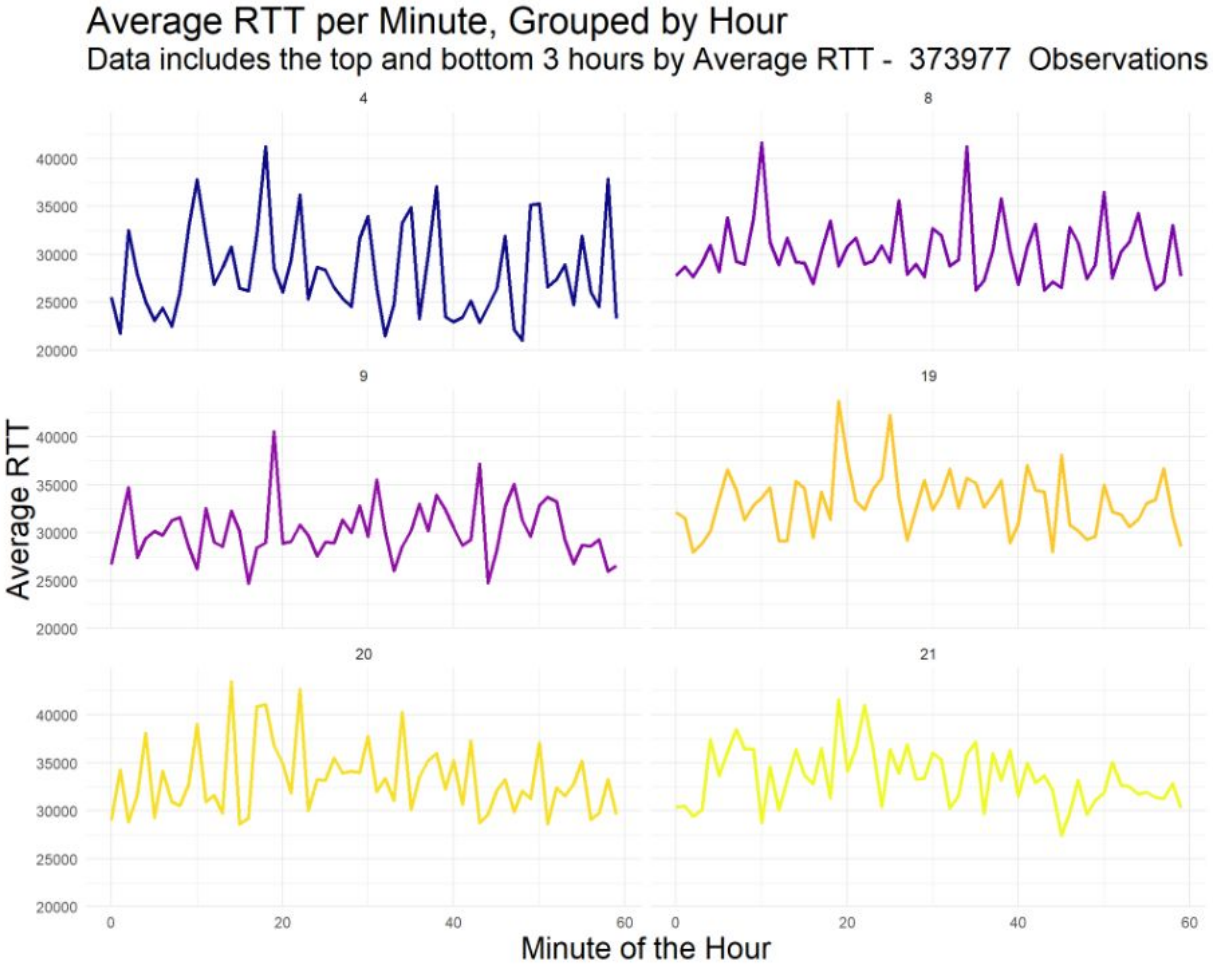


Please note that when we refer to time of day, we refer to the local daylight savings time. We will separate this topic into the following sub-questions:

1. Does DNS performance differ depending on time of day?

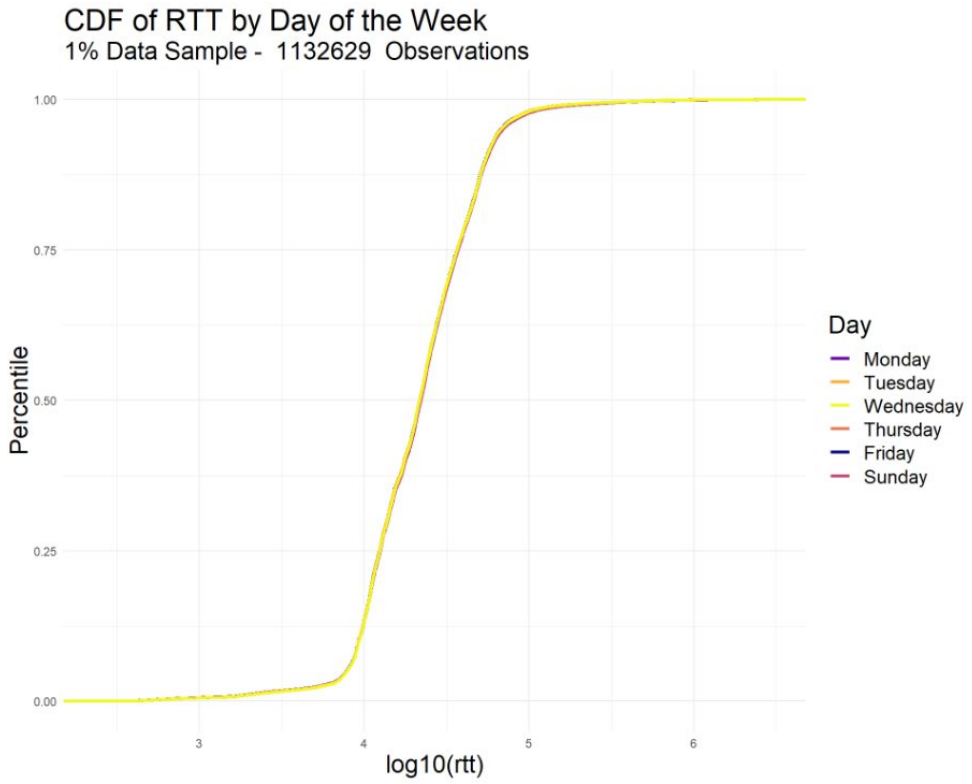
We predict that latency will increase with internet usage throughout the day, with periods of high internet usage producing high rtt's and vice versa. The resulting data confirms this prediction, but to less of an extent than expected. The general trend follows general internet usage trends over the course of a day, which confirms that higher web traffic leads a decrease in DNS performance. However, average rtt ranges from roughly 28500 microseconds at the low end to roughly 33000, a ~16% increase. While significant, we expected a far more drastic effect, which speaks positively of the servers' abilities to handle large loads. We conclude that DNS performance does differ depending on time of day, with the best and worst time of days producing rtt's that differ by <20%.


Average RTT per Hour
1% of Data Sample - 1134921 Observations
Average RTT
32000
30000
28000
5
10
15
20
Hour Since Midnight

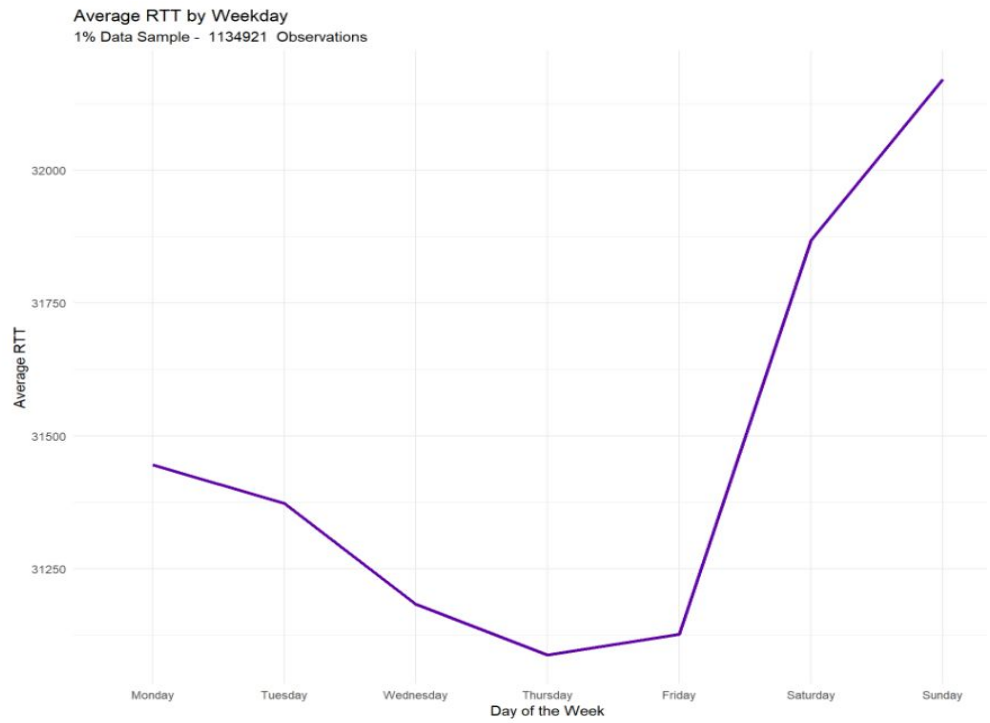

Average RTT per Minute, Grouped by Hour
Data includes the top and bottom 3 hours by Average RTT - 373977 Observations
4
8
9
19
20
21
Average RTT
40000
35000
30000
25000
20000
0
20
40
60
Minute of the Hour

2. Does DNS performance differ depending on day of week?

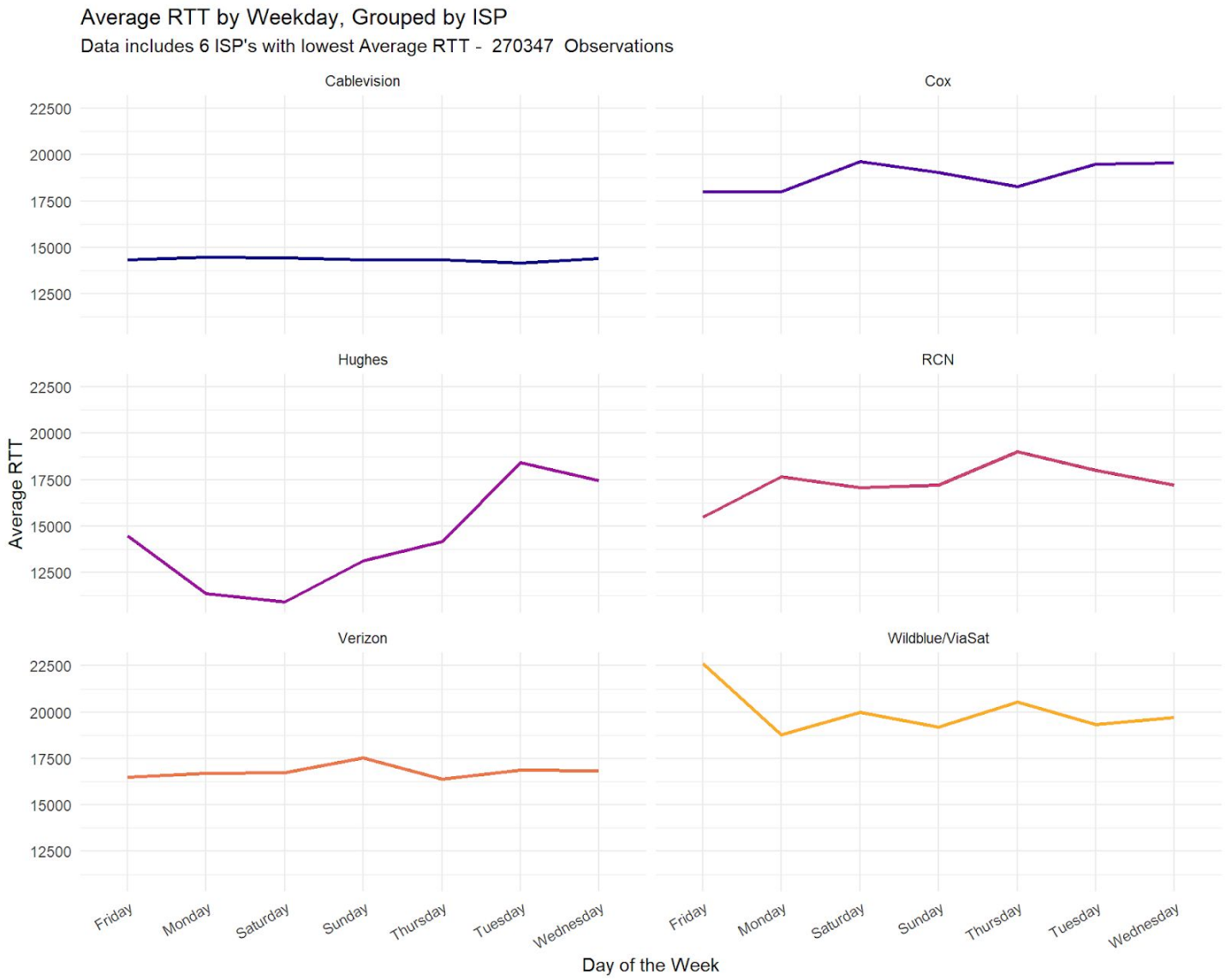
We expect that DNS resolution times correlate with web traffic, with a high volume of web traffic increasing the latency of DNS responses. As such, we predict higher latency on Fridays through Sundays, where people have more leisure time for web browsing, streaming, gaming, and other web intensive tasks and activities. The below chart shows the CDFs of rtt separated by the day of the week on which the tests were issued:



Surprisingly, we see negligible differences among the CDFs.



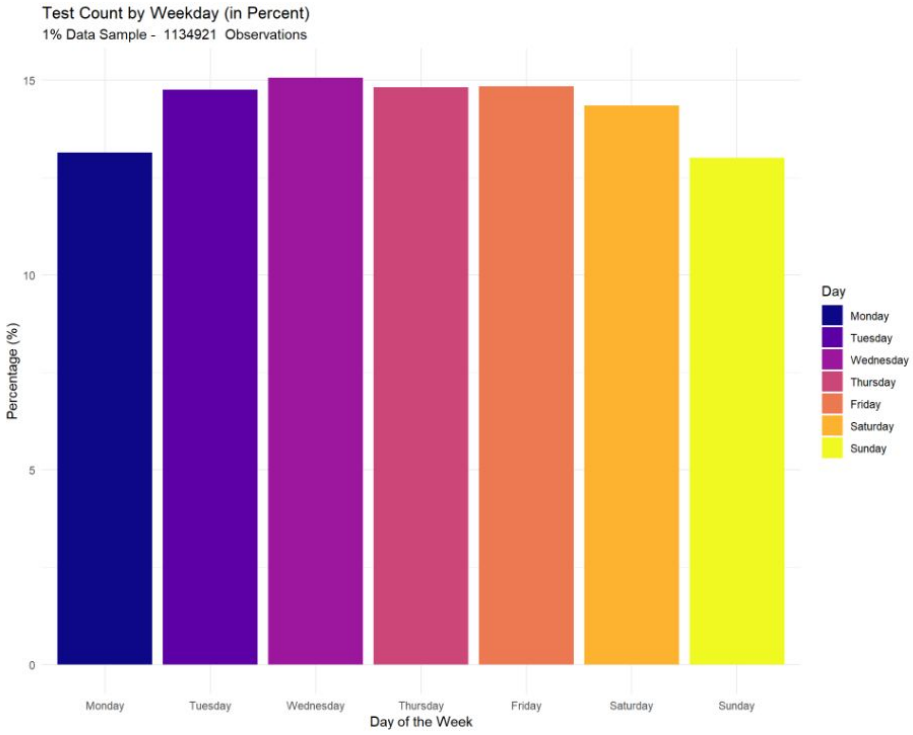
The above graph, showing average RTT across different days of the week, gives a better visualization of differences across different weekdays. As expected, weekends produce the highest RTT's, but we are surprised to see that RTT falls throughout the workweek. However, we note that the scale of the graph exaggerates the differences seen across the different days of the week, and that RTT values only range from ~30750 microseconds to ~31750 microseconds, which is arguably a negligible difference. To further explore this, we plot the same graph, but grouped by ISP to see if the difference in performance still remains negligible throughout the week for different internet providers by choosing 6 ISPs with the lowest average RTT .



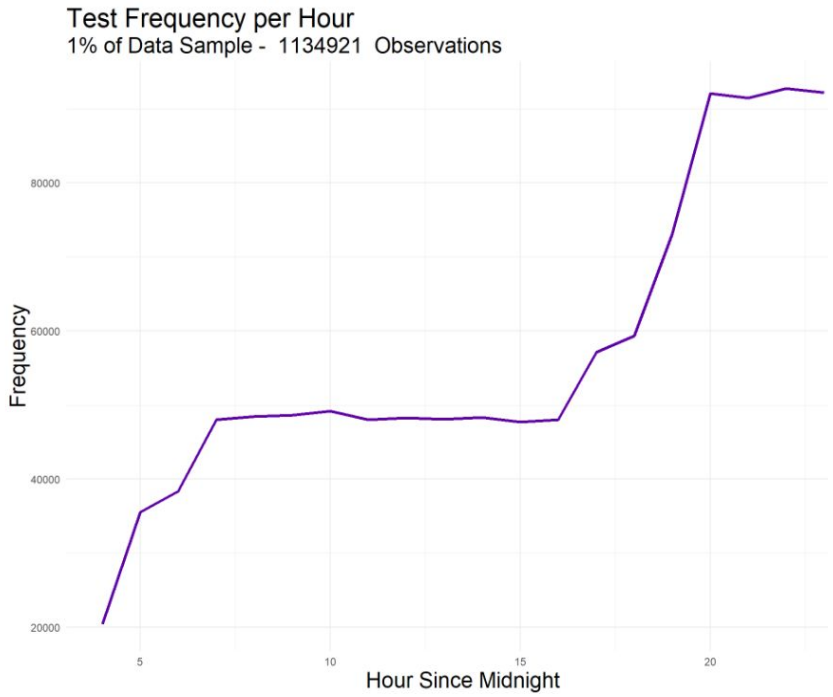
From these plots we can see that average RTT varies greatly by ISP and there does not seem to be a common trend between these 6 ISPs. For example, Hughes maintains a consistent ~14,000 average RTT per day, while larger companies like Verizon range from roughly 12000 to 18000 within a week. From this, DNS performance is more affected by provider than by day of the week, but overall performance by day seems negligible.

3. Does the FCC test different ISP's at different times?

We would like to test whether the FCC is consistent in conducting their tests under similar conditions across time, so that test statistics do not unfairly over- or under-represent certain times of day or days of the week.

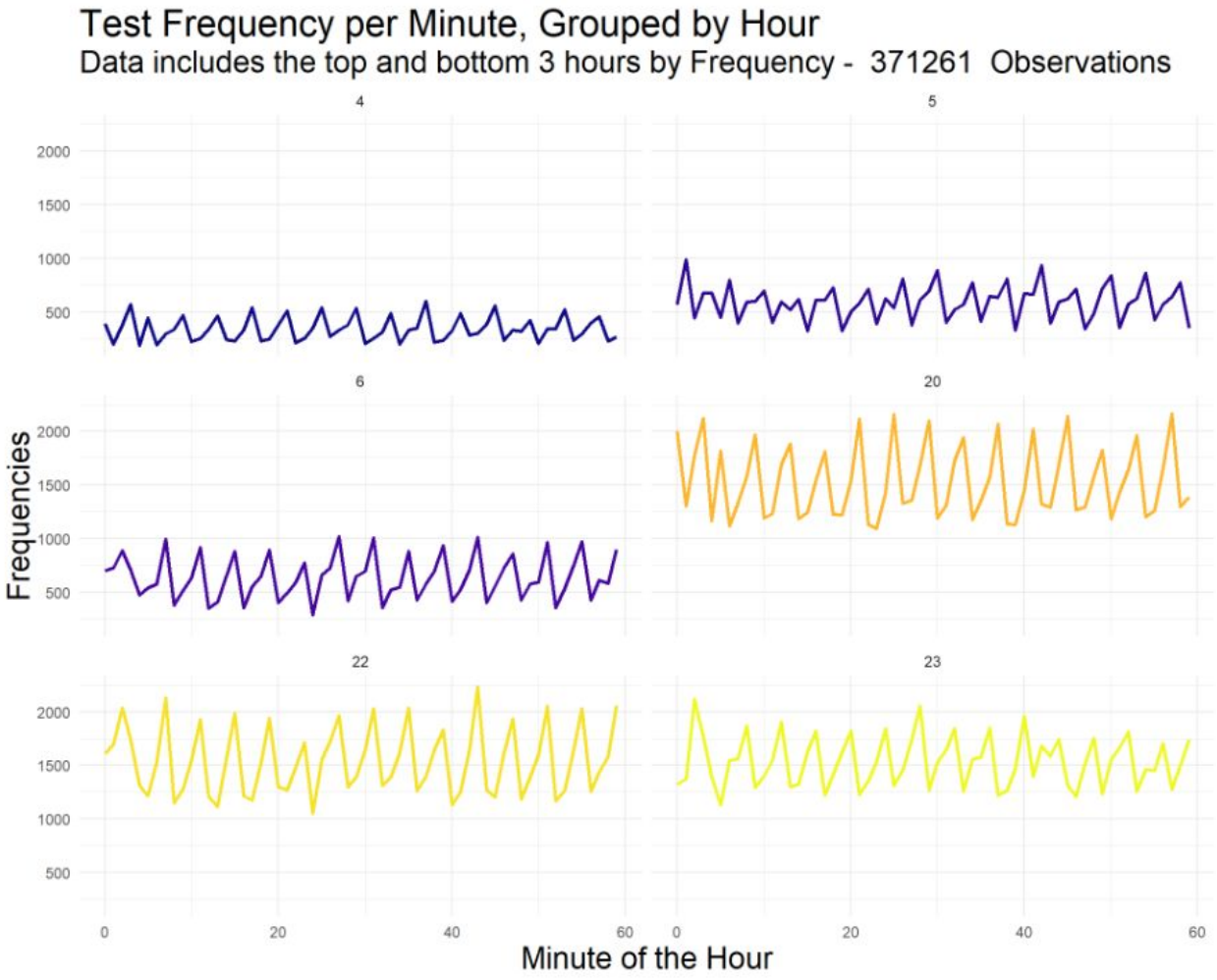


As shown in the chart above, the FCC seems to issue fewer tests on Sundays and Mondays, with very little difference in the number of tests issued on the other days of the week. As a result, any test statistics we use will slightly under-represent any data collected from Sunday and Monday. However, the difference is small enough that we do not feel these differences will have a significant impact on our conclusions.

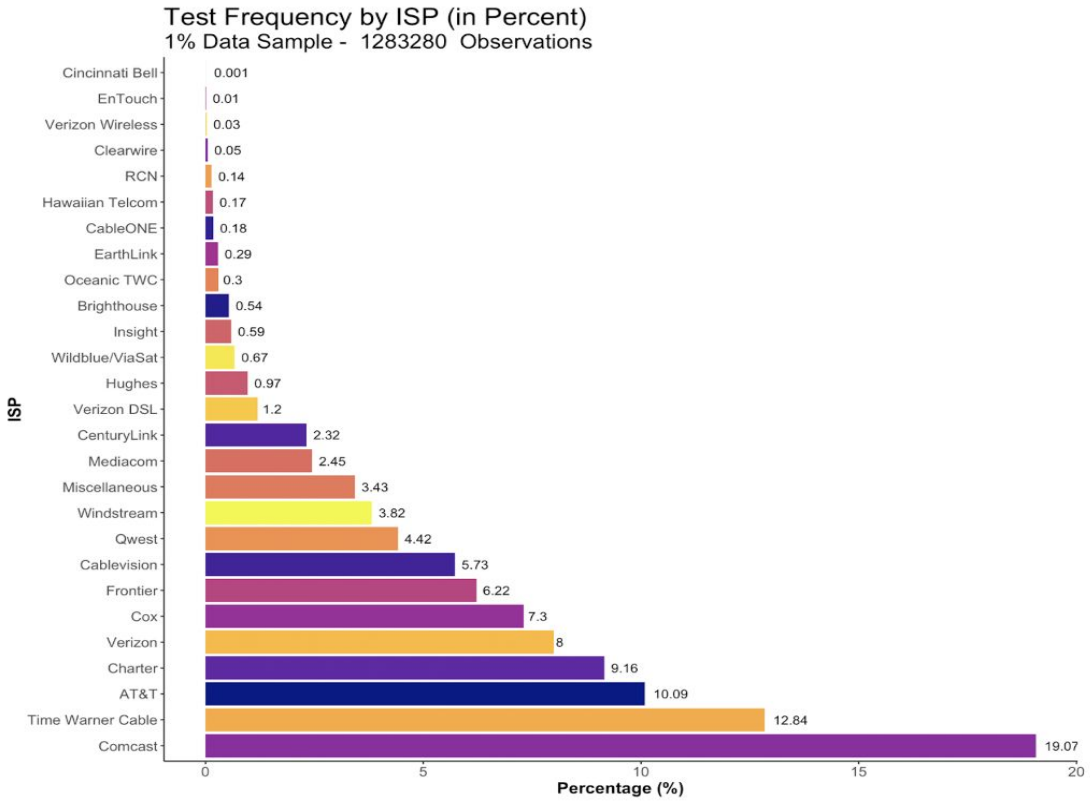


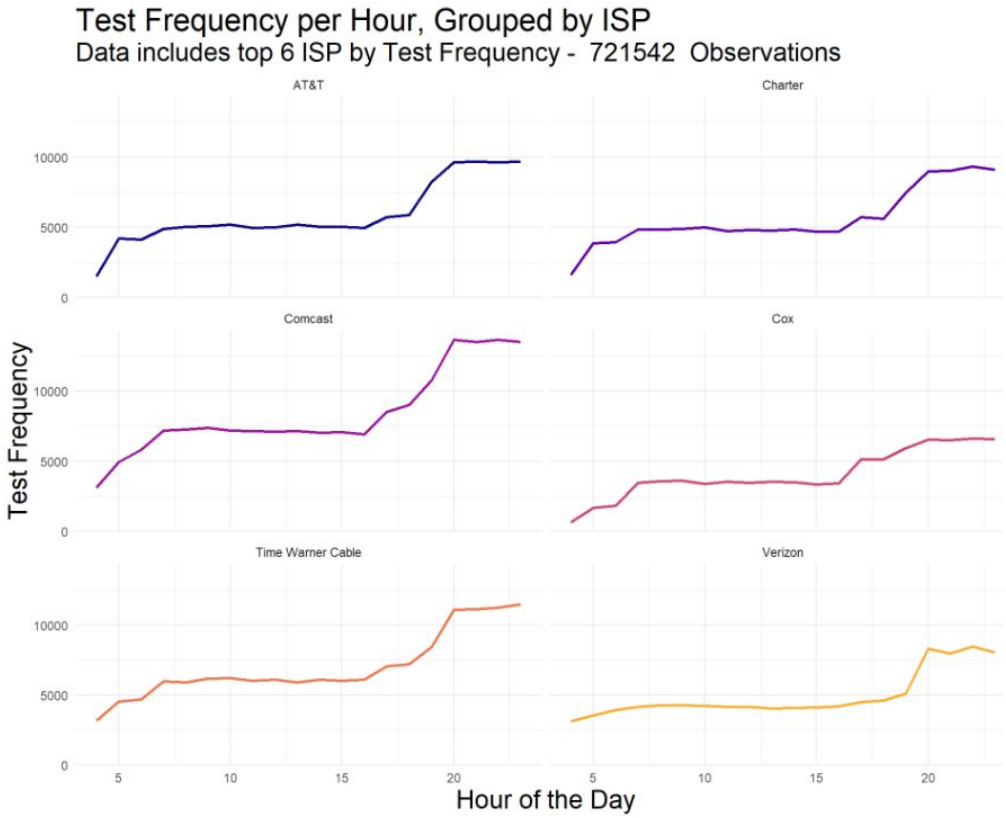
However, we do see variation in the number of tests the FCC issues throughout the day. We see that the queries begin daily at 4AM and continue to rise in volume throughout the day until they peak at 8PM. This pattern mirrors the average American's use of the internet shown in the beginning of Part 2. This suggests that the FCC issues tests throughout the day with frequency rising with internet usage. This would suggest that the FCC likely issues tests to ensure DNS performance when networks are under heavy load from peak internet usage. The lack of tests between midnight and 4AM could either reflect FCC employees' work schedules or could suggest that the FCC does not care about DNS performance during times of low internet usage, likely because it can assume low-latency performance. Either way, our results will not indicate DNS performance in the early hours of the morning.

If we look further into the data and examine tests over the course of a single hour grouped by minute, we see a curious pattern of the number of tests spiking every 4 minutes. We cannot think of a concrete explanation for this phenomenon, but this frequency could reflect limitations in the FCC's technology or could simply indicate an arbitrarily decided frequency of tests.



Now that we have gained some insight into the overall test frequencies by time of day/day of week, we will further analyze this with respect to ISP. Looking at the bar chart below we see the breakdown of test frequencies by ISP. We will choose to examine the top 6 ISPs based on test frequency in the said chart (Comcast, Time Warner Cable, AT&T, Charter, Verizon, Cox) due to them having the most influence on the overall test frequency time series above.





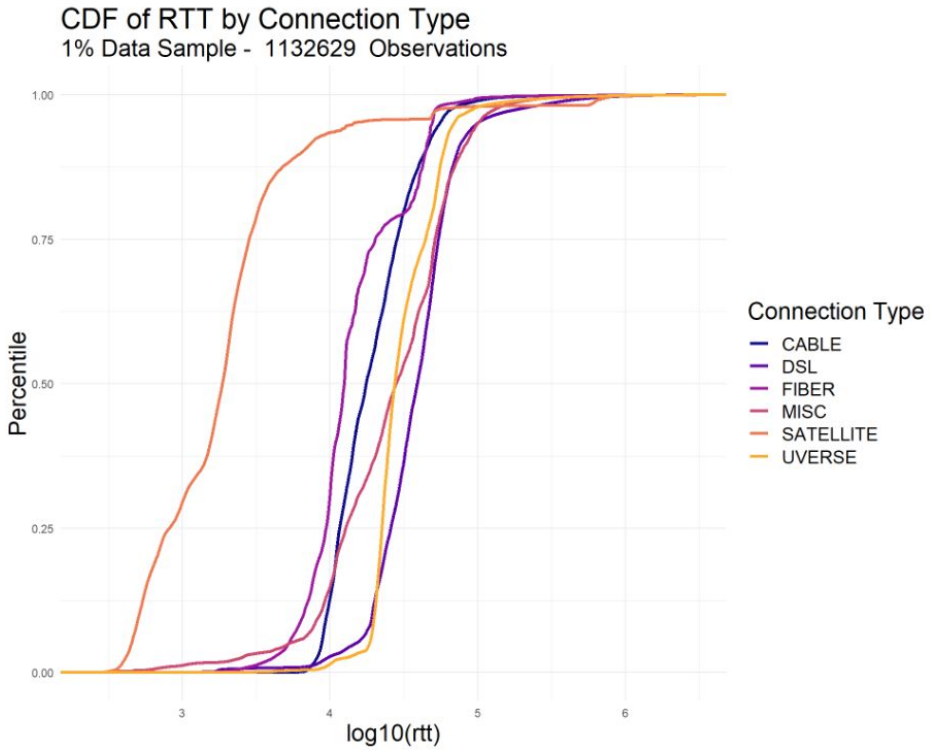
In the figure above, we have gathered time series of test frequencies per hour of the day, grouped by the top 6 ISPs. One thing to note is that the magnitude of the test percentage for each ISP closely aligns with how similar it mirrors the overall test count time series. For example, Comcast, which has the highest percentage, looks almost identical to the overall time series while Cox, which has the lowest percentage, looks the least identical. However, all these time series resemble each other and, to varying degrees, the overall test count time series. It is unlikely that the FCC tests different ISPs at different times.

PART III

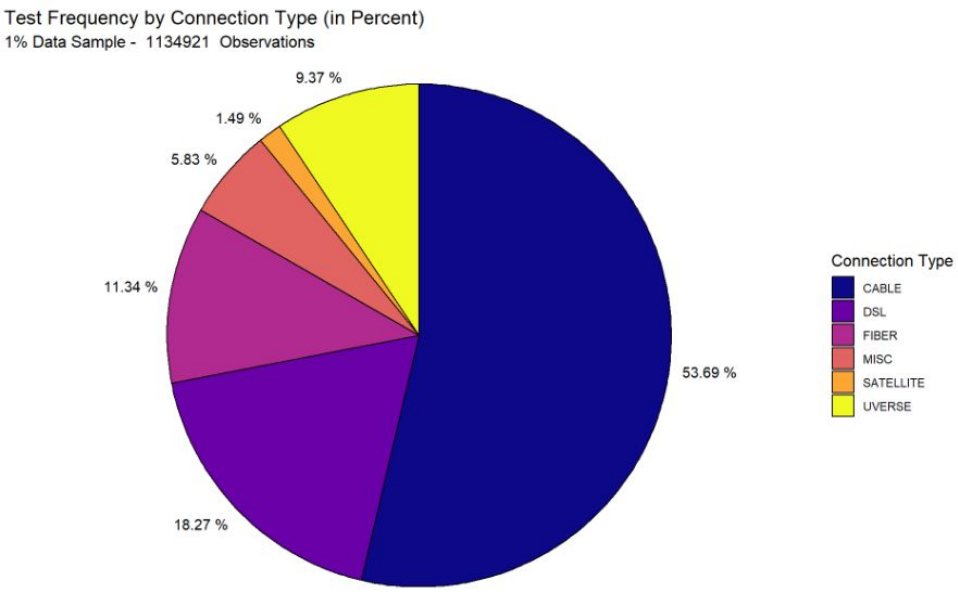
How does DNS performance differ by ISP and connection type pair?

Perhaps one of the most influential factors in determining DNS performance may connection type, as it reflects differences in technology and infrastructure employed by the local area and the ISP. While newer technologies such as fiber optics may produce vastly improved performance, limited resources, limited population, or the presence of geographical boundaries may hinder an area's ability to install such infrastructure. We examine the effects of these differing connection types and ISPs here.

1. How does connection type influence DNS performance?

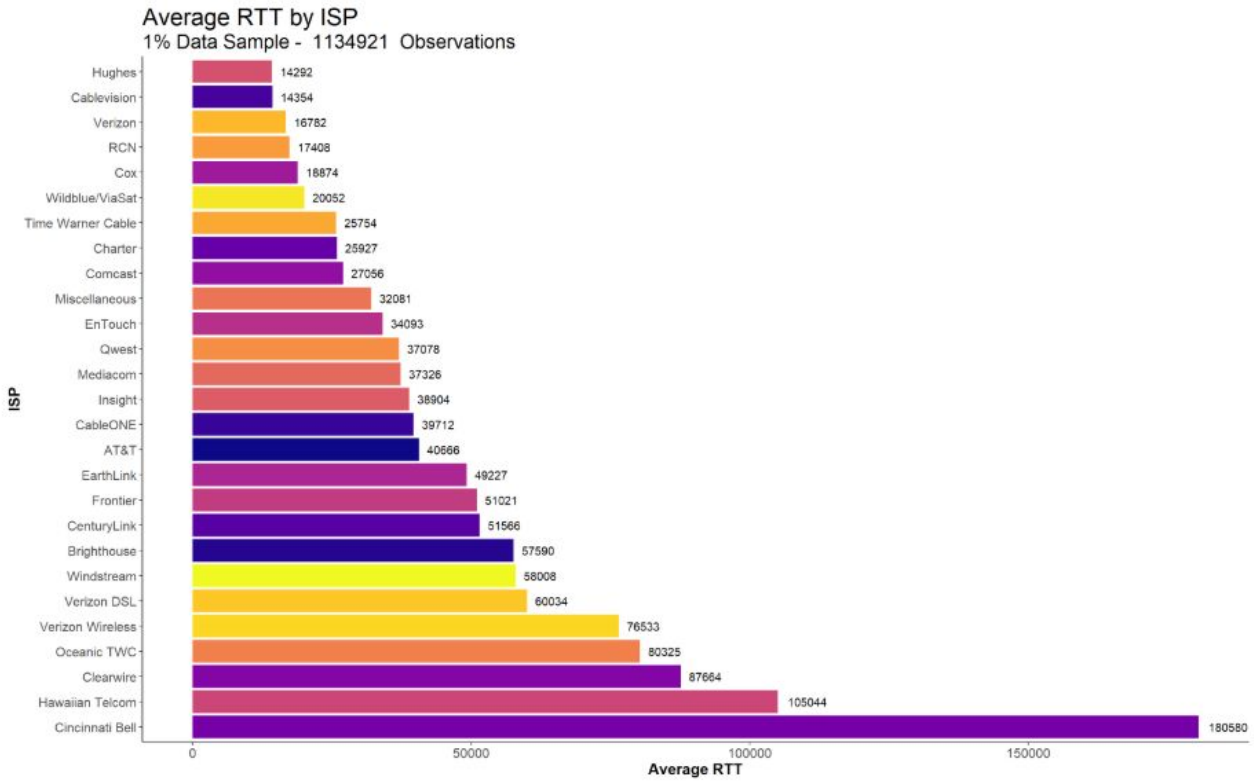


The above chart shows the stark differences between rtt's when utilizing different connection types. In particular, satellite connections produce significantly smaller rtt's with fiber in distant second, while DSL connections produce larger rtt's. Satellite's position with the lowest rtt's in the dataset is an abnormality not only because its CDF is so distant from the others, but also because of its status as one of the slowest connection types out of modern day internet access options. This can be explained by the fact that data for satellite is abnormal in that it is both sparse and may not be representative of the technology as a whole. Satellite data represents little more than 1% of the data while more than 60% of the satellite data is represented by a single ISP. Based on previous knowledge and the abnormalities present in the graph, satellite data is essentially an outlier so we will continue to examine our cdfs without it. Since we present rtt on a logarithmic scale, the increase in the median rtt from roughly ~10^4 microseconds with fiber connections to ~10^4.5 microseconds with DSL connections represents a roughly threefold increase in latency. This emphasizes the inequity in modern day internet access. While DSL uses existing telephone lines and are therefore widely available, fiber optic cables are a relatively new technology and require their own separate installation process, which can be expensive. This could, in effect, widen the disparities between DNS performance in wealthy or populous areas and poor or sparsely populated areas.



Luckily, a mere 16% of tested connections were on DSL, while almost half of all connections were on cable, which performs significantly better than DSL. Additionally, according to current trends, DSL is continuing to decline while fiber connections continue to grow. So, while a sizeable portion of the U.S. population still uses slow DSL connections, that segment of the population is slowly moving towards better and faster technologies.

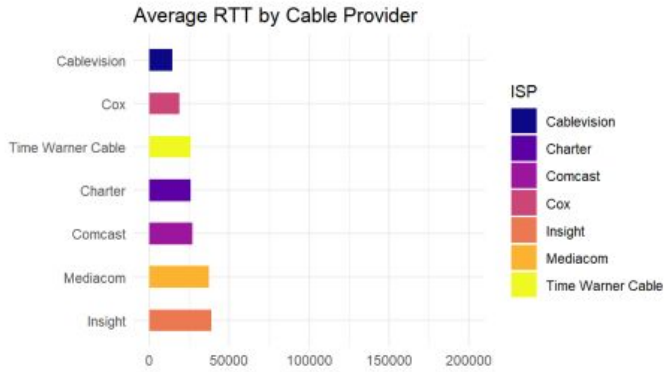
2. How does ISP influence DNS performance?

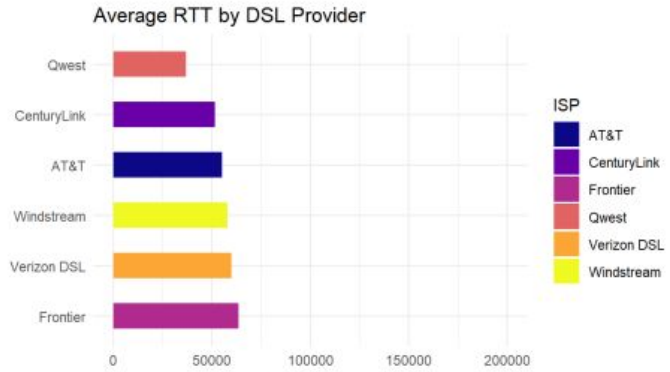


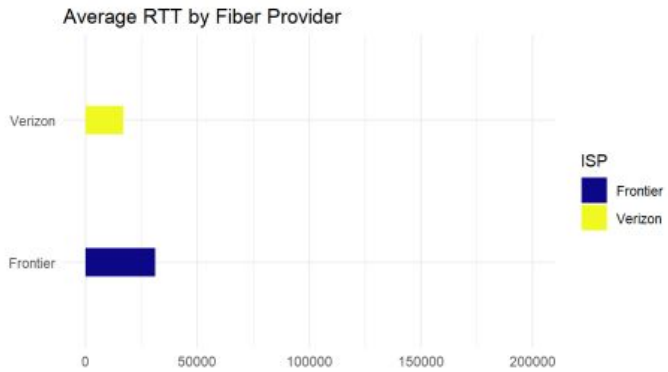
The above graph shows the average rtt's of different ISPs. Cincinnati Bell comes in distant last in terms of performance, while Hughes performs the best.
The table to the right shows all the unique combinations of ISP and connection type present in the dataset. To better visualize this table, the 6 graphs below show the average RTT of ISPs grouped by connection type. These grouped figures show that connection type does seem to affect the average RTT of each group. However, since only AT&T and Frontier utilize more than one connection type, we will not investigate ISP and connection type simultaneously, as it will be equivalent to our evaluation of just one.

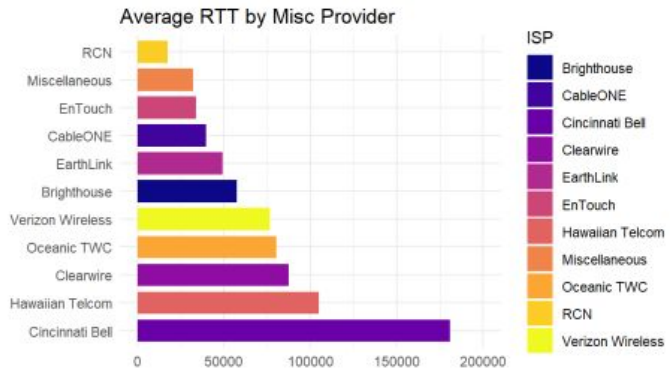
ISP <fctr>	Connection Type <fctr>	Test Count <int>
AT&T	DSL	12687
AT&T	UVERSE	106334
Brighthouse	MISC	6946
CableONE	MISC	2324
Cablevision	CABLE	70882
CenturyLink	DSL	29574
Charter	CABLE	113191
Cincinnati Bell	MISC	123
Clearwire	MISC	661
Comcast	CABLE	168106
Cox	CABLE	81452
EarthLink	MISC	3764
EnTouch	MISC	143
Frontier	DSL	46683
Frontier	FIBER	29441
Hawaiian Telcom	MISC	2193
Hughes	SATELLITE	10159
Insight	CABLE	7561
Mediacom	CABLE	27645
Miscellaneous	MISC	44016
Oceanic TWC	MISC	3799
Qwest	DSL	55443
RCN	MISC	1820
Time Warner Cable	CABLE	140535
Verizon	FIBER	99237
Verizon DSL	DSL	15368
Verizon Wireless	MISC	407
Wildblue/ViaSat	SATELLITE	6797
Windstream	DSL	47630

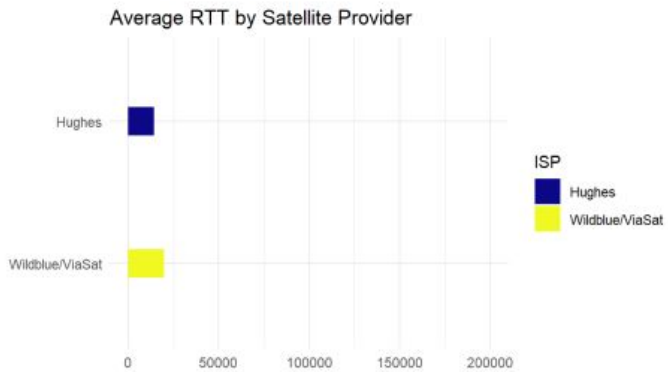
29 rows

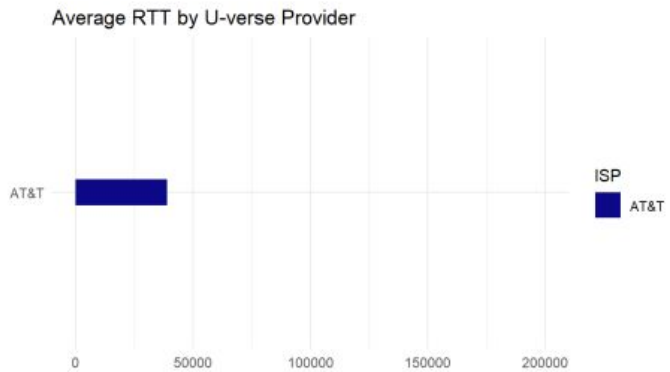












CONCLUSION

We now conclude our report by recapping the major findings of our report:

- Black box distribution correlates highly with population of each state. This suggests that black boxes are randomly assigned to individuals or households, rather than locations.
- The Northeast has significantly better DNS performance than the other 3 regions.
- States with high population densities tend to have better DNS performance and vice versa.
- Rtt's throughout a given day roughly follow the trends of general internet usage throughout the day, with periods of high volume of internet traffic having high rtt's
- There is little difference among average rtt's on different days of the week
- The FCC issues tests according to general internet usage trends, issuing more tests in periods of high internet usage. The FCC does not issue tests between midnight and 4AM.
- The FCC is approximately fair in testing different ISPs.
- Satellite performs best overall as a connection type, but we question the data's validity on this point, since so little of the data consists of a satellite connection type. Fiber comes out in second, while DNS is the poorest performer.
- Cincinnati Bell had the poorest DNS performance out of all ISPs, while Hughes had the best DNS performance
- Most ISPs utilize only a single connection type.