\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Memorandum**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

To: Derek Wu, Neil Cholli

From: Dylan Craig

Date Created: 12/20/2024

Date Updated: 12/21/2024

Subject: Data Analysis Memo on Office Distances (CommonHelp VA)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Purpose:** To provide descriptive statistics on the results of adding Haversine distances, driving distances/time, and transit distances and time between each VDSS local office zip code and the midpoints of each zip code associated with that local office (using Google API).

**Part 1: Data Collection**

Data was collected using historical screenshots (via Wayback Machine) of the [VDSS local offices website](https://www.dss.virginia.gov/localagency/index.cgi) containing addresses and dates addresses were used. The HUD USPS Zip Code Crosswalk was used to determine which zip codes were associated with which VDSS local offices based on county(ies) served. The final unit of observation is office address/date/associated zipcode.

Coordinates for the VDSS local office as well as the midpoint of each zip code served at least partially by that local office were calculated using Google API with some manual corrections made using Google Maps for obvious outliers.

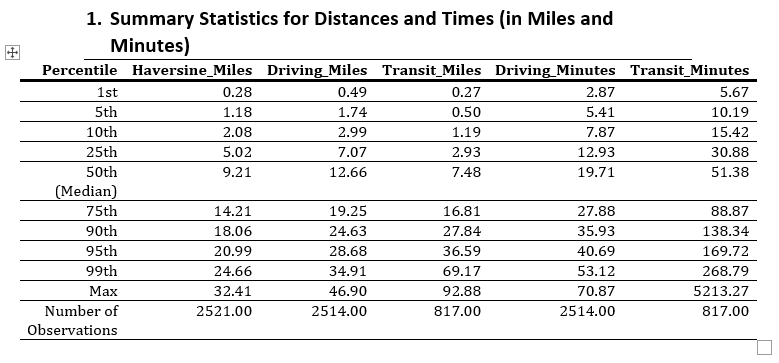
Then distance measures were calculated between each local office and their associated zip codes including the following: Haversine Distance (“as the crow flies”) (meters), Driving Distance (meters), Driving Time (seconds), Transit Distance (meters), Transit Time (seconds). Calculation was completed again using Google API.

**Part 2: Data Analysis**

*Data Cleaning:*

After loading the data, some initial cleaning was done. First, prior to analysis, meters were converted to miles and seconds converted to minutes. Secondly, duplicate instances of LDSS zip code coordinates and associated zip code coordinates were reduced, leaving 3012 unique observations out of 3833 original observations.

*Summary Statistics:*

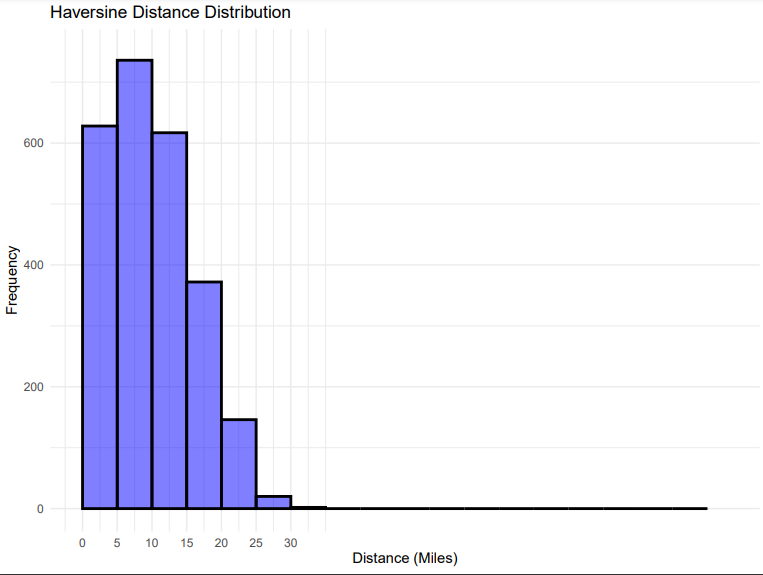
**

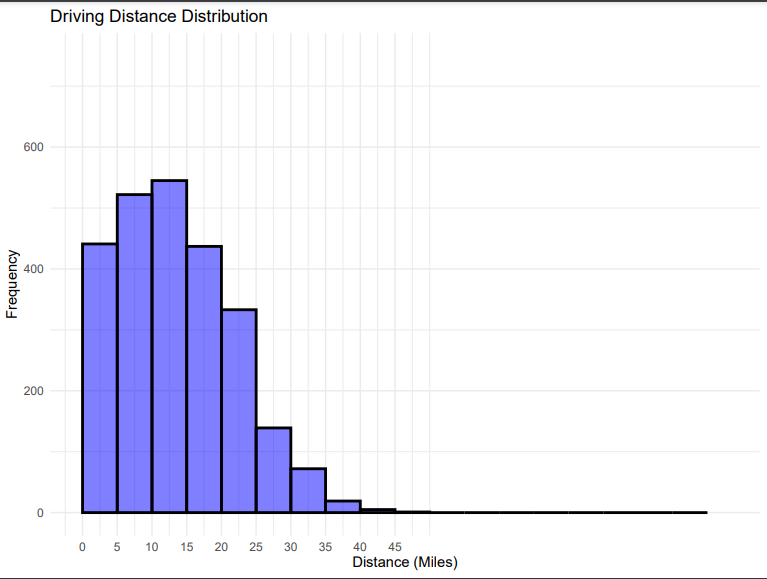
To start, summary statistics for each distance were calculated at various percentiles. These percentiles were calculated independently for variables, irrespective of if the row had observations for each mode of travel (haversine, driving, transit). Transit had far fewer observations than Haversine distances, and its observations were clustered around VDSS local office - zip pairs that were shorter in distance (see missing Transit data section). Therefore, comparing Transit data with Haversine or Driving data based on these percentiles would be inaccurate for stating whether Transit distance is generally “longer or shorter” than haversine or driving distance.

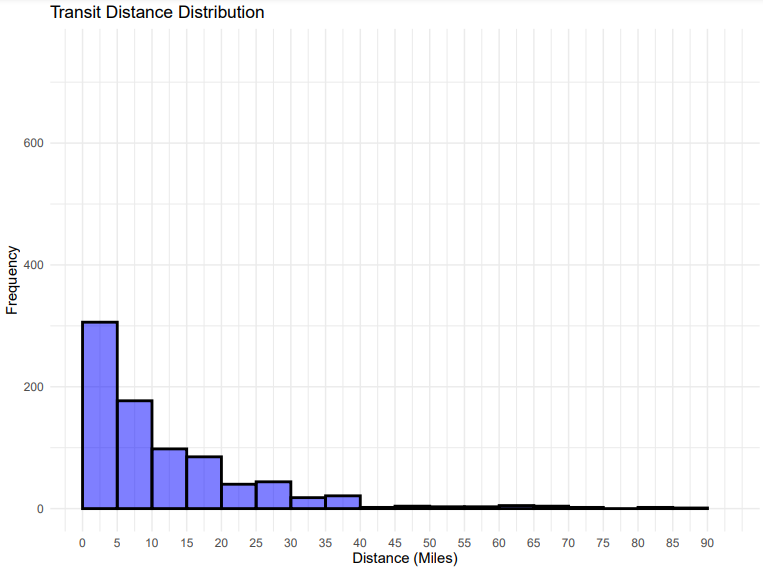
However, haversine distance and driving distance share approximately the same number of observations and we can infer that driving distance generally is longer than haversine distance, as expected.

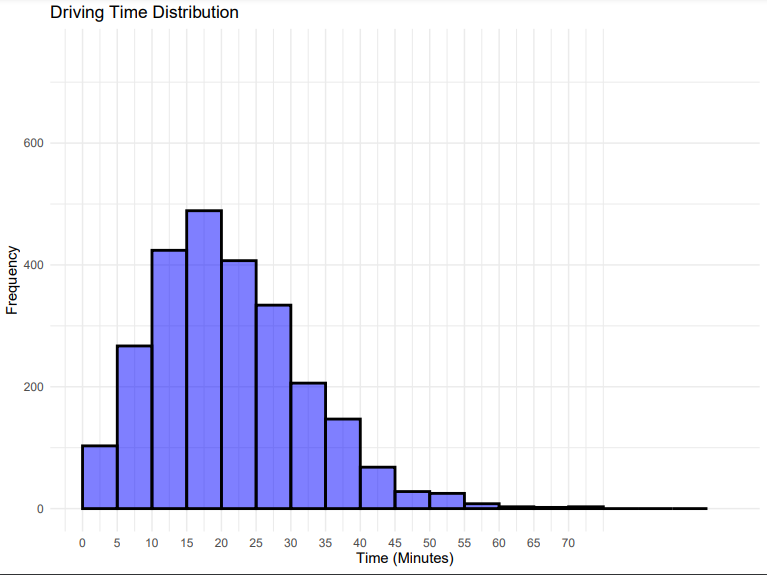
*Distribution of Distances (Histogram)*

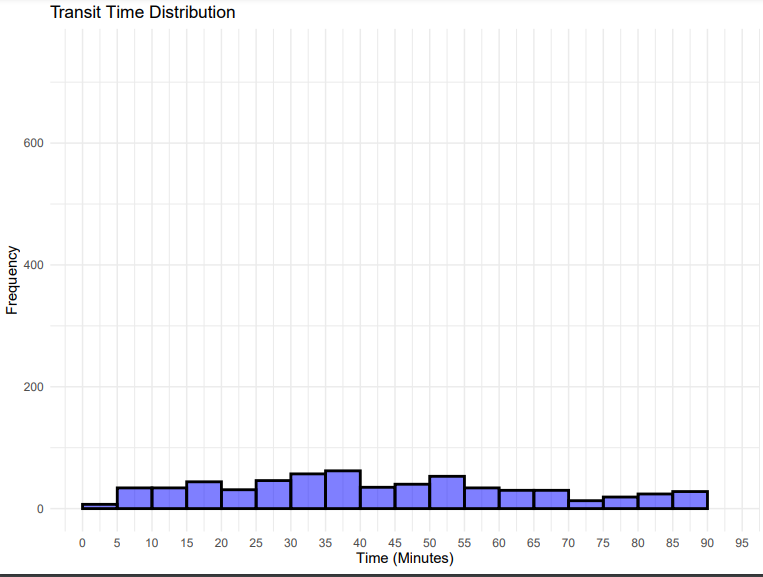
To visualize the frequency distribution of different distances and travel times, several histograms were created. X and Y axes were held fixed to ensure accurate comparison.







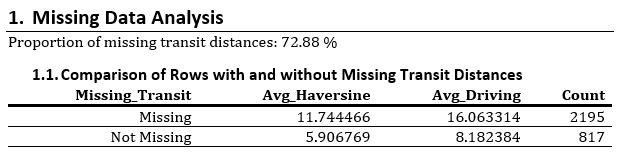




*Missing Transit Distances:*

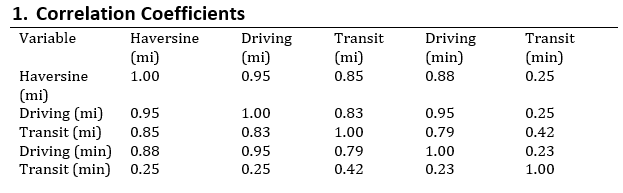
In order to better understand the missing transit data in the sample, data was broken up into Transit = Missing and Transit != Missing.

72.88% of the sample was missing transit data and observations with missing transit data showed, on average, significantly higher Haversine and Driving Distances. This makes sense, given that Google is less likely to identify a transit option for more rural areas or areas with greater distances between a zip code midpoint and its associated VDSS office.



*Correlation Table:*

Finally, a correlation table was created between our 5 different measurements of travel distance. To be careful, only rows with complete observations (so haversine, driving, and transit) were included.



The table shows strong correlations (above 0.8) between Haversine distance, driving distance, and driving time, which makes sense since these measures are closely related to road networks and straight-line distances. Transit distance also has a strong correlation with Haversine and driving distances but slightly weaker, reflecting how transit routes can deviate from direct paths. Transit time, however, has weak correlations with other variables (around 0.25-0.42), likely because of the variability in public transit systems like stops, transfers, and delays. The strongest relationship overall is between driving distance and driving time (0.95), which is expected due to consistent travel speeds. These findings highlight how road-based metrics align more closely, while transit measures are less predictable.

Again, these findings are likely concentrated around observations with lower haversine distances, and may look different if we were to include observations with higher haversine distances.

*Haversine Ventile Analysis*

# Haversine Ventile Analysis

| HAVERSINE\_VENTILE | Avg Haversine Miles | Avg Driving Miles | Avg Transit Miles | Avg Driving Minutes | Avg Transit Minutes | Share Missing Driving Miles (%) | Share Missing Transit Miles (%) | Share Missing Driving Minutes (%) | Share Missing Transit Minutes (%) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 0.61 | 1.00 | 0.97 | 4.14 | 17.98 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 | 1.61 | 2.52 | 2.74 | 7.85 | 36.13 | 0.00 | 10.24 | 0.00 | 10.24 |
| 3 | 2.58 | 3.86 | 5.64 | 9.37 | 53.51 | 0.00 | 36.51 | 0.00 | 36.51 |
| 4 | 3.58 | 5.32 | 6.85 | 12.02 | 49.41 | 0.00 | 49.21 | 0.00 | 49.21 |
| 5 | 4.59 | 6.55 | 9.70 | 13.03 | 77.32 | 0.00 | 51.59 | 0.00 | 51.59 |
| 6 | 5.48 | 7.98 | 10.68 | 14.70 | 75.59 | 0.00 | 65.08 | 0.00 | 65.08 |
| 7 | 6.33 | 9.01 | 15.39 | 16.50 | 107.14 | 0.00 | 66.67 | 0.00 | 66.67 |
| 8 | 7.08 | 10.11 | 15.26 | 17.57 | 110.94 | 0.00 | 76.98 | 0.00 | 76.98 |
| 9 | 7.98 | 11.45 | 16.67 | 19.11 | 121.34 | 0.79 | 74.60 | 0.79 | 74.60 |
| 10 | 8.80 | 12.58 | 20.19 | 20.18 | 131.35 | 0.00 | 75.40 | 0.00 | 75.40 |
| 11 | 9.65 | 13.00 | 21.58 | 19.60 | 149.47 | 0.00 | 77.78 | 0.00 | 77.78 |
| 12 | 10.60 | 14.54 | 17.53 | 21.79 | 93.69 | 0.00 | 81.75 | 0.00 | 81.75 |
| 13 | 11.50 | 15.77 | 21.18 | 23.52 | 107.67 | 0.00 | 79.37 | 0.00 | 79.37 |
| 14 | 12.50 | 17.00 | 22.87 | 25.27 | 130.66 | 0.00 | 85.71 | 0.00 | 85.71 |
| 15 | 13.69 | 18.56 | 28.42 | 27.50 | 206.25 | 0.00 | 87.30 | 0.00 | 87.30 |
| 16 | 14.73 | 19.94 | 25.36 | 28.97 | 165.55 | 0.79 | 84.92 | 0.79 | 84.92 |
| 17 | 15.79 | 21.77 | 31.66 | 31.20 | 137.48 | 1.59 | 91.27 | 1.59 | 91.27 |
| 18 | 17.13 | 22.91 | 41.65 | 32.55 | 215.79 | 0.00 | 84.13 | 0.00 | 84.13 |
| 19 | 19.44 | 25.62 | 47.85 | 36.30 | 347.51 | 0.79 | 89.68 | 0.79 | 89.68 |
| 20 | 23.49 | 30.98 | 54.26 | 42.22 | 142.01 | 0.00 | 91.27 | 0.00 | 91.27 |
| NA | NaN | NaN | NaN | NaN | NaN | 100.00 | 100.00 | 100.00 | 100.00 |

The table breaks down average haversine miles, driving miles, transit miles, driving minutes, transit minutes, as well as the percent of missing values for driving miles, transit miles, driving minutes, and transit minutes by each haversine distance ventile.

A few conclusions can be made:

1. On average, across all but the first ventile, the order of nearest to furthest distance is as follows: haversine distance, driving distance, transit distance
2. The gap between transit and other forms of distance, both in miles and minutes, grows pretty exponentially as haversine distance increases
   1. This is likely due to the unreliability of transit options as distance increases
3. The share of missing driving values is practically zero across all ventiles whereas the percent of missing transit values increases significantly after the 1st ventile.