Bruce Ramseyer
06/19/23

Immersion Task 6.1 – Sourcing Open Data

Data Source:

Trips by Distance	
Source	This is an external data source from the Department of Transportation via Data.gov (https://catalog.data.gov/dataset/trips-by-distance). The data is considered trustworthy since it has been provided by a government agency.
Collection	Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory estimated the data for Bureau of Transportation Statistics based off an anonymized national panel of mobile device data from various sources. The data was first published on 8/17/2021 with the most recent update as of 2/1/23. A weighting procedure was used to expand the sample to make it representative of the entire US population.
Contents	Information in this dataset includes collection date, geographical level category, state, county, population of individuals regarding their travel status, total trips based off defined distance categories, and time-dependent variables featuring week and month travel totals. The data contains 4,867,656 records in 22 rows and has been aggregated at the national, state, and county level.
Limitations	The initial merged mobile device data had to meet certain temporal frequency and spatial accuracy as part of the project. The Bureau of Transportation Statistics has indicated that the data is experimental and may not meet all their quality standards. The information has been published to the benefit of data users where no other data has been available. The Bureau of Transportation Statistics is seeking validation to assist in regular publication.
Ethics	The data has been anonymized and aggregated to eliminate any reference of the mobile device user. Applications of this data do not seem predisposed to any sort of bias. An argument could be made that the government should not be collecting data from individuals' mobile devices. If the mobile carriers eliminated any sensitive personal information prior to sharing the data, this would eliminate the concern.
Relevance	The information within the dataset has great significance as collection began prior to the COVID-19 outbreak and has been continuously updated throughout the pandemic. Analysis of the data will provide insights into how much COVID impacted travel within the United States and whether individuals have returned to pre-COVID levels.

Bruce Ramseyer
06/19/23

Data Profile:

The data consistency checks were performed in Python. Mixed data types were updated to string values. No duplicate values were found. Null values were present at State, County and Trip Totals with patterns present within each category.

State and County null values were a result of National and State categories levels. The National level does not contain any records for State and County. State level does not contain any records for County. As such no additional actions were needed on the 1524 State and 79248 null values.

The remaining 39424 null values at a County level were further analyzed. Null values were present in most states. Alaska, Texas, and Montana have the highest total counts, all of which have remote/rural county features. Alaska was further reviewed and a comparison between the most and least populous boroughs was performed. The most populous showed zero no values whereas the least populous borough had the highest null values for the Alaska subset. Seasonality also appeared to be a factor for Alaska null values, but this trend did not hold true for every state.

The County null values on the continuous variables were ultimately dropped as they represented only 0.82% of the total data and most likely featured remote and/or less populous counties amongst the entire dataset. Furthermore, the null values may have been anonymized as part of data collection as the Bureau of Transportation Statistics stated, "no data are (sic) reported for a county if it has fewer than 50 devices in the sample on any given day," to assure confidentiality and support data quality.

Variable	**Description**	**time - variant/ -invariant**	**structured/ unstructured**	**qualitative/ quantitative**	**qualitative: nominal/ ordinal quantitative: discrete/ continuous**
Level	Category of aggregates at a National, State, and County level	invariant	structured	qualitative	ordinal
Date	Date of data collection	variant	structured	qualitative	nominal
State FIPS	The Federal Information Processing Standards code for state level classification	invariant	structured	qualitative	nominal
State Postal Code	State abbreviation as used by US Postal Service	invariant	structured	qualitative	nominal
County FIPS	The Federal Information Processing Standards code for county level classification	invariant	structured	qualitative	nominal
County Name	Name of the specific county within each state	invariant	structured	qualitative	nominal

Population Staying at Home	Total number of individuals that did not take any trips	invariant	structured	quantitative	continuous
Population Not Staying at Home	Total number of individuals that did take a trip of any length	invariant	structured	quantitative	continuous
Number of Trips	Total number of trips recorded on the collection date	invariant	structured	quantitative	continuous
Number of Trips <1	Total number of trips recorded that were less than a mile	invariant	structured	quantitative	continuous
Number of Trips 1-3	Total number of trips recorded that were between 1 to 3 miles	invariant	structured	quantitative	continuous
Number of Trips 3-5	Total number of trips recorded that were between 3 to 5 miles	invariant	structured	quantitative	continuous
Number of Trips 5-10	Total number of trips recorded that were between 5 to 10 miles	invariant	structured	quantitative	continuous
Number of Trips 10-25	Total number of trips recorded that were between 10 to 25 miles	invariant	structured	quantitative	continuous
Number of Trips 25-50	Total number of trips recorded that were between 25 to 50 miles	invariant	structured	quantitative	continuous
Number of Trips 50-100	Total number of trips recorded that were between 50 to 100 miles	invariant	structured	quantitative	continuous
Number of Trips 100-250	Total number of trips recorded that were between 100 to 250 miles	invariant	structured	quantitative	continuous
Number of Trips 250-500	Total number of trips recorded that were between 250 to 500 miles	invariant	structured	quantitative	continuous
Number of Trips >=500	Total number of trips recorded that were greater than or equal to 500 miles	invariant	structured	quantitative	continuous
Row ID	Unique identifier for each row within the dataset	invariant	structured	qualitative	nominal
Week	The week of the total aggregated trips	variant	structured	qualitative	nominal
Month	The month of the total aggregated trips	variant	structured	qualitative	nominal

Bruce Ramseyer
06/19/23

Data Exploration:

1. How does the amount of travel vary between urban and rural counties?
2. How does mobility, individuals staying at home versus those who do not, vary between state and county?
3. How has the number of trips changed before, during, and after the COVID-19 pandemic?
4. Which states were the most and least travel impacted by COVID-19?
5. What lengths of trips were most impacted by COVID-19?