Influenza Influences: Prediction of Influenza Vaccine Distribution Using Search Engine Query Data

G2 - The InFLUenzers

Data Science Capstone Project
Exploratory Data Analytics Report

Date:
May 27, 2022

Team Members:
Name: Ashley Wheeler
Name: Andrew Chen
Name: Tien Nguyen
Name: Jackie Glosson

Our Approach to Handling Missing Data

1. Data with missing values that we interpolated:
    a. WHITE, BLACK, HISPANIC, RACE_OTHER, PERCENT_ATTAIN_BACHELORS, MEDIAN_HOUSEHOLD_INCOME
    b. We had six columns with missing data, of which we tested out two different methods of interpolation. In our interpolation, we utilized two approaches (Approach 1 and Approach 2) described in detail in section 3. We compare the differences between these approaches in this Exploratory Data Analysis Report to determine which may be the best approach. However, we run through this EDA using Approach 1 as our final dataset for simplicity's sake. Be aware that the only difference between Approach 1 and 2 and our data with missing variables comes from these six columns.
    c. PERCENT_OVER_65,
2. Data with missing values we did not interpolate, but used repetitive values:
    a. AVG_TEMP, PERCENT_VOTED_DEM, PERCENT_VOTED_REPUBLICAN, PERCENT_UNINSURED, PERCENT_PRIVATE_INSURED, PERCENT_PUBLIC_INSURED
3. Data not missing any values:
    a. GET FLU SHOT, FLU SHOT, FLU SHOT NEAR ME, FLU VACCINE, FLU SHOT SIDE EFFECTS, ESTIMATE %

Analysis of the primary metrics of variables

Categorical or continuous: All variables except for STNAME_MONTH are numerical. STNAME_MONTH is our only categorical feature composed of the state name, calendar month, and year. Refer to Table 1 for the list of attributes and their meanings.

Table 1. Attributes and Their Meanings

Attribute	Data Type	Number of NA	Meaning of Attribute
GET FLU SHOT	float	0	Google Trend search term
FLU SHOT	float	0	Google Trend search term
FLU SHOT NEAR ME	float	0	Google Trend search term
FLU VACCINE	float	0	Google Trend search term
FLU SHOT SIDE EFFECTS	float	0	Google Trend search term
PERCENT_OVER_65	float	560	Percent of people who are 65-year-old and above (%)
VAX_PERCENT_DISTRIBUTION	float	469	Percent of people getting vaccinated (%)
MEDIAN_HOUSEHOLD_INCOME	float	580	Median household income in dollars ($)

PERCENT_ATTAIN_BACHELORS	float	580	Percent of people who are 25-year-old and have bachelor's degrees or higher (%)
WHITE	float	570	Percent of people who identify as white (%)
BLACK	float	570	Percent of people who identify as black (%)
RACE_OTHER	float	570	Percent of people who identify as other races (%)
HISPANIC	float	590	Percent of people who identify as Hispanic (%)
AVG_TEMP	float	0	Average temperature (^{o}F)
PERCENT_VOTED_DEMOCRATIC	float	0	Percent of people who voted for demographic (%)
PERCENT_VOTED_REPUBLICAN	float	0	Percent of people who voted republican (%)
PERCENT_UNINSURED	float	0	Percent of people who do not have insurance (%)
PERCENT_PRIVATE_INSURANCE	float	0	Percent of people who have private insurance (%)
PERCENT_PUBLIC_INSURANCE	float	0	Percent of people who have public insurance (%)

Basic metrics of data without missing values: Note that our original google trends search term "FLU SHOT" has notably the largest standard deviation (or largest variance) of all search terms. We can know why this is by learning how the google trends data is obtained. Google trends data normalizes data with respect to the queries (maximum of five) examined together. In this instance, the query of "FLU SHOT" was searched more times than any other query overall – we know this because google trends assigns a maximum value of 100 to the data point representing the most significant number of searches for that period of time and group of queries. We can also detect that the term "FLU SHOT" was searched the most from the mean. When we consider which search query to use in our model, we will want to take this into account, as the other search queries may not capture as much variance as our FLU SHOT query since the way it is normalized through google trends "compresses" the other queries.

The google trends search term "FLU SHOT" has notably the largest standard deviation (or largest variance) of all search terms. We can know why this is by learning how the google trends data is obtained. Google trends data normalizes data with respect to the queries (maximum of five) examined together. In this instance, the query of "FLU SHOT" was searched more times than any other query overall – we know this because google trends assigns a maximum value of 100 to the data point representing the most significant number of searches for that period of time and group of queries. We can also detect that the term "FLU SHOT" was searched the most from the mean. When we consider which search query to use in our model, we will want to take this into account, as the other search queries may not capture as much variance as our FLU SHOT query since the way it is normalized through google trends "compresses" the other queries. Therefore we will select the "FLU SHOT" variable for our google trends variable.

Basic data metrics prior to handling missing values (See Appendix Figure B): Overall, we see minimal variance (small standard deviation) in our demographic variables. Minimal variance is expected, as these values within states change very little over time- perhaps only a few decimal places. For example, PERCENT_OVER_65, WHITE, BLACK, PERCENT_VOTED_DEMOCRATIC, and PERCENT_VOTED_REPUBLICAN all have standard deviations of less than one. This is also expected given the missing data. We will compare to see if interpolation changes the variance of these columns.

Comparison of data before and after handling missing values (See Appendix Figure C and D): We do not observe large differences in the standard deviations when comparing values interpolated in approach 1 versus Approach 2, as opposed to with missing values. In fact, all standard deviations of demographic variables are within 0.1 standard deviations of each other, meaning that our interpolation has not drastically altered the composition of our data. The most considerable difference in standard deviation is seen in MEDIAN_HOUSEHOLD_INCOME. Our original standard deviation was 13,754. Approach 1 had a standard deviation of 13,630, and Approach 2 had a standard deviation of 14,399. A similar trend is observed for the column's median. In this instance, Approach 1 seems closer to our original data than Approach 2. Both approaches had very similar standard deviations to our original data for the variable PERCENT_ATTAIN_BACHELORS. Initial data had a standard deviation of 7.05, Approach 1 had a standard deviation of 6.97, and Approach 2 had a standard deviation of 6.98. In summary, our basic metrics suggest that our interpolation approaches have upheld the composition of the original data, with a preference for Approach 1.

Non-graphical and graphical univariate analysis

A. Distribution before and after missing data is handled

Histograms of data without missing values: We observe that all google trends data is right-skewed, and Vax_Percent_Distribution (our target feature) is similarly right-skewed. Likewise, PERCENT_OVER_65 is slightly right-skewed, and AVG_TEMP is slightly left-skewed.

In most columns where missing data was repeated rather than interpolated, we observe some abnormally shaped distributions simply because there are repetitive values. For example, our voting data columns (PERCENT_VOTED_DEMOCRAT and PERCENT_VOTED_REPUBLICAN) have odd shapes. Each column has two values per state: One for percent voting in 2016 and one for percent voting in 2020. We repeat the exact data for all rows of each state, leading to histograms that look far from normally distributed and instead have a relatively "blocky" appearance. The same can be said for the three insurance columns (private, public, and other) and the column PERCENT_OVER_65, in which similar strategies to fill in missing data were employed. All of these have an abnormal, "blocky" distribution.

Comparison of distribution before and after handling missing values (See below Figure 1 and Appendix E and F): Examining the distributions of the racial columns, we observe that both WHITE and BLACK have binomial distributions in all three variations of data: the data with missing values, as well as both approaches. The other racial columns all have the "blocky" abnormal distributions, similar to the columns in which we repeated data rather than interpolated.

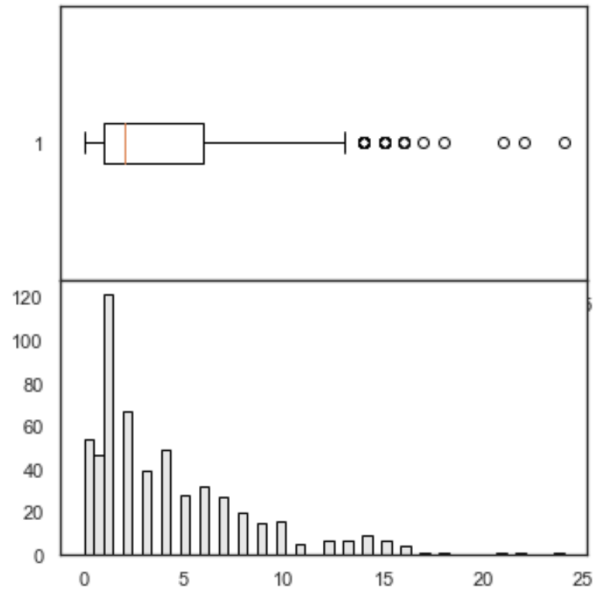
Examining the distribution of the variable PERCENT_ATTAINED_BACHELORS, we observe that the distributions for all three histograms look similar. For MEDIAN_HOUSEHOLD_INCOME, we observe that the histogram for Approach 1 and our original data look very similar, while Approach 2 has a spike previously unobserved in the other histograms.

Figure 2. Histograms of Numerical Variables after data pre-processing with Approach 1:

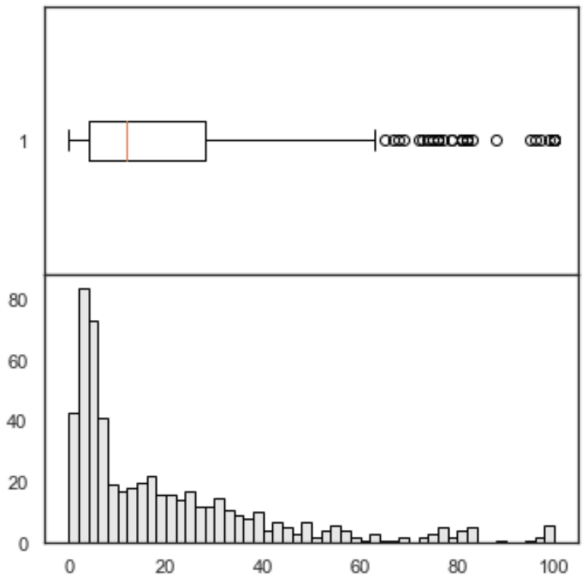
B. Boxplots

Visualizing the spread of our numerical independent variables and their outliers:

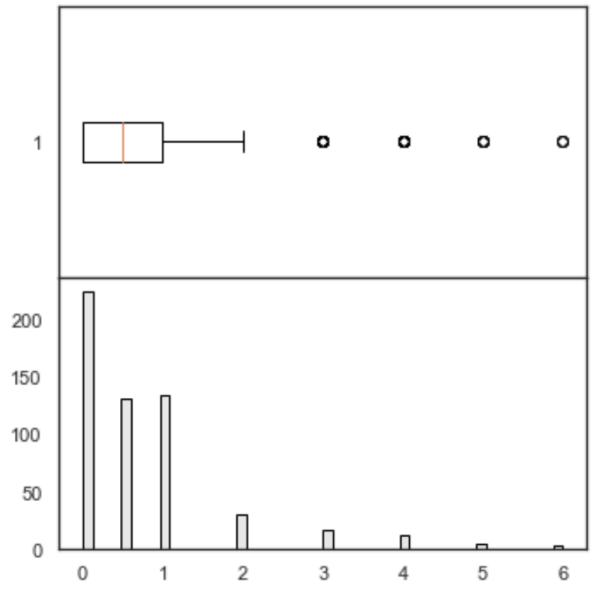
GET FLU SHOT



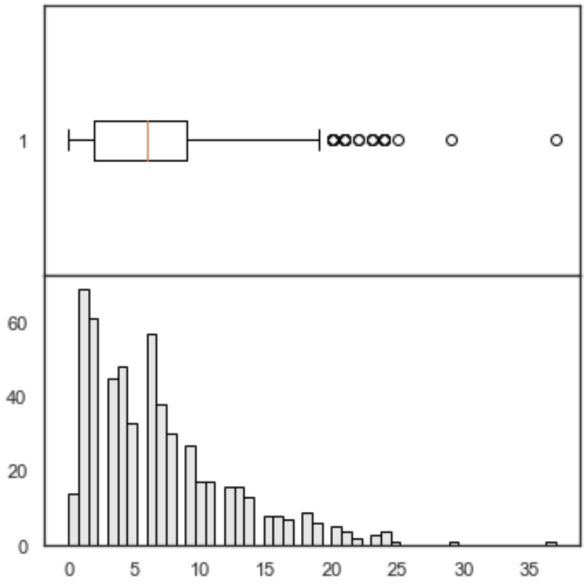
FLU SHOT



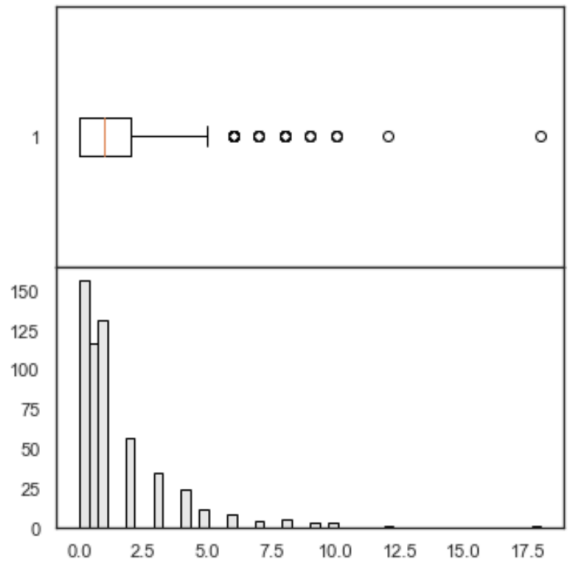
FLU SHOT NEAR ME



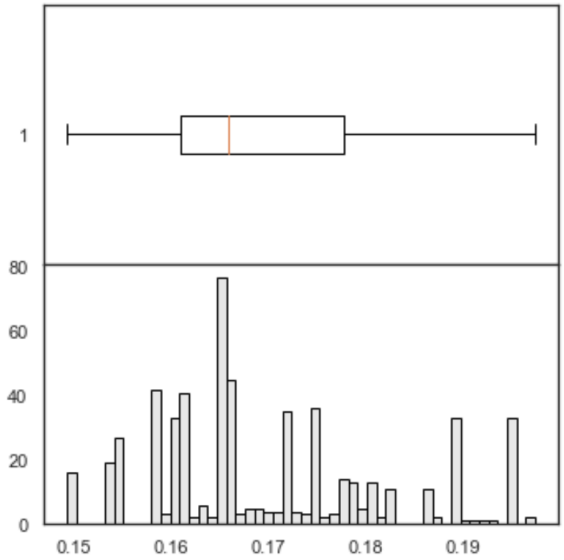
FLU VACCINE



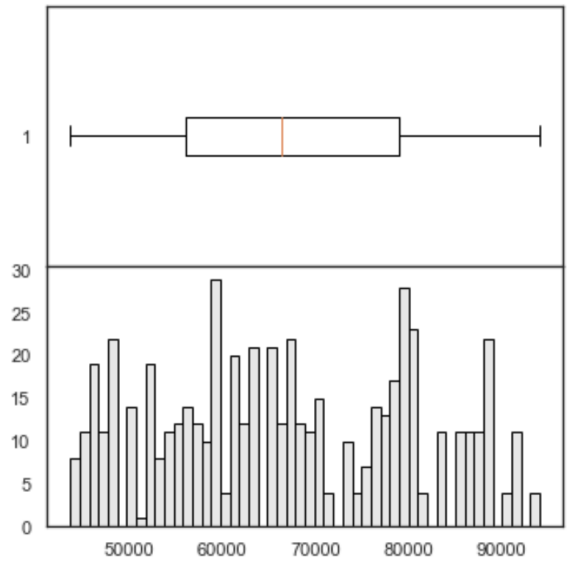
FLU SHOT SIDE EFFECTS



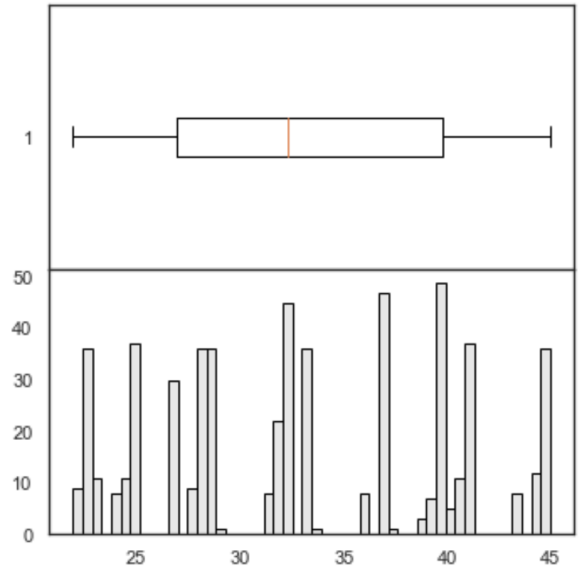
PERCENT_OVER_65



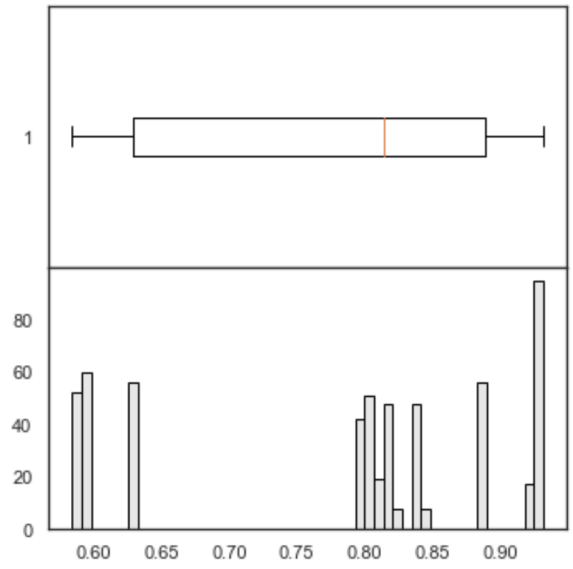
MEDIAN_HOUSEHOLD_INCOME



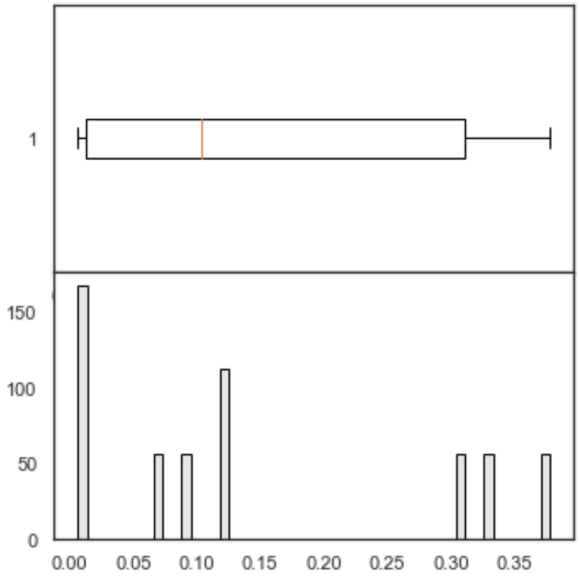
PERCENT_ATTAIN_BACHELORS



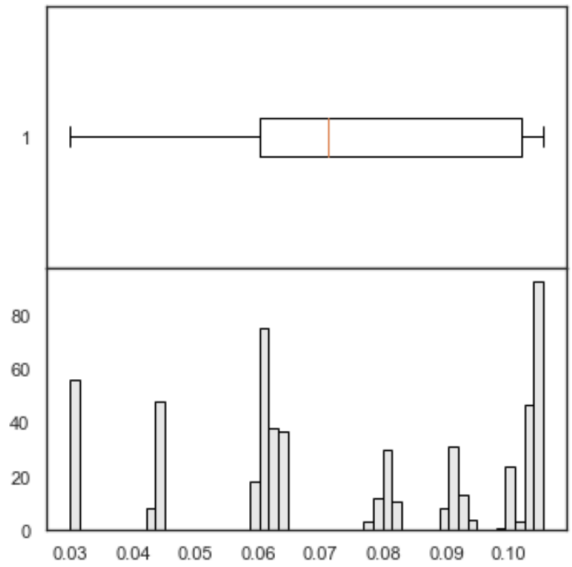
WHITE



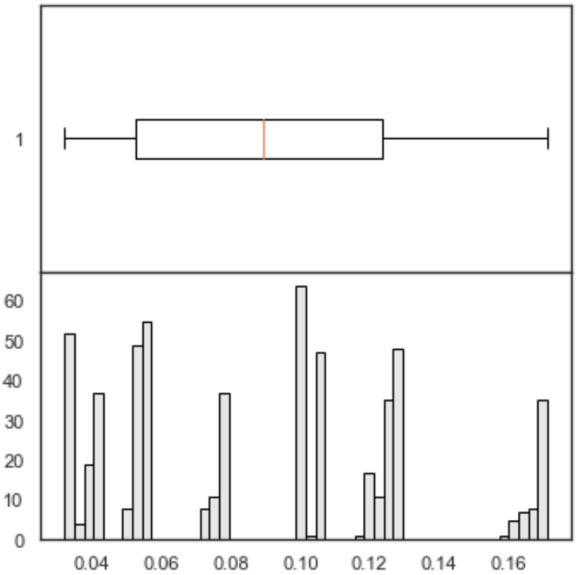
BLACK



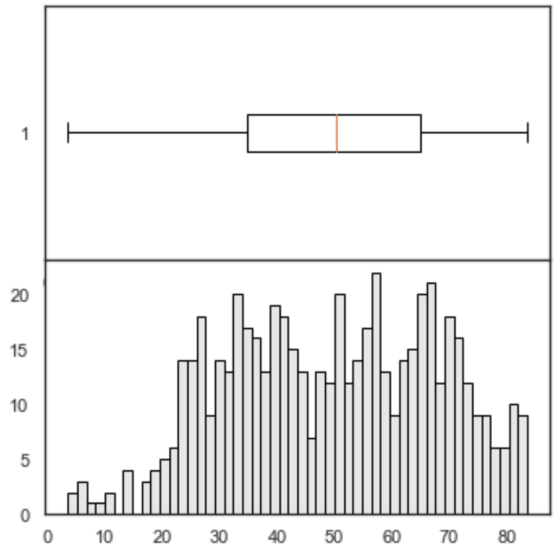
RACE_OTHER



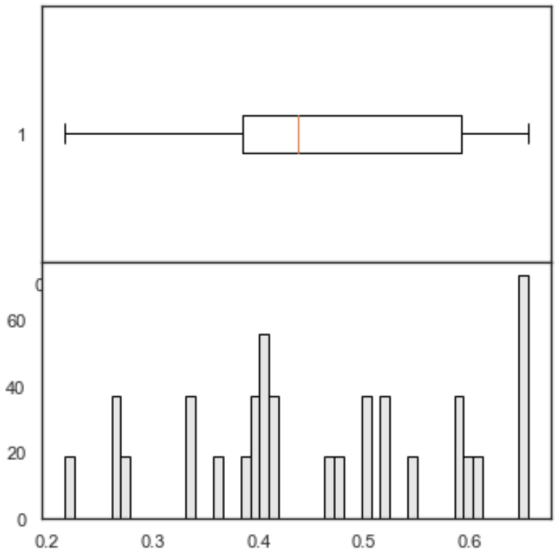
HISPANIC



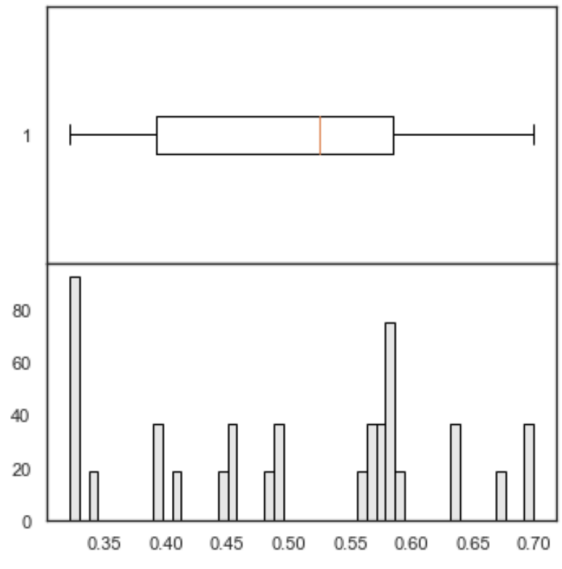
AVG_TEMP



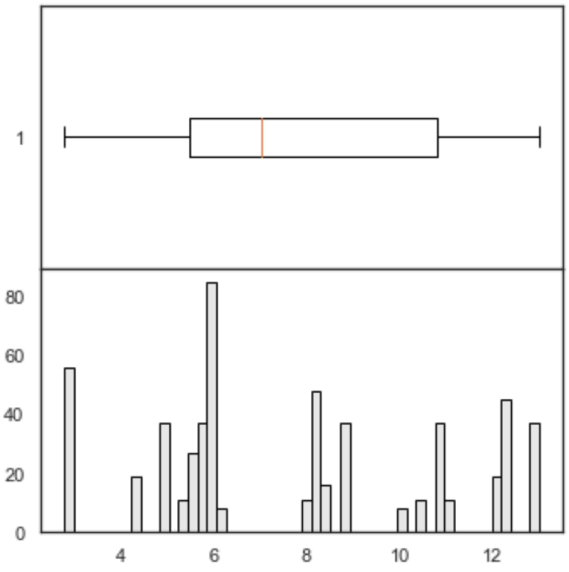
PERCENT_VOTED_DEMOCRATIC



PERCENT_VOTED_REPUBLICAN



PERCENT_UNINSURED



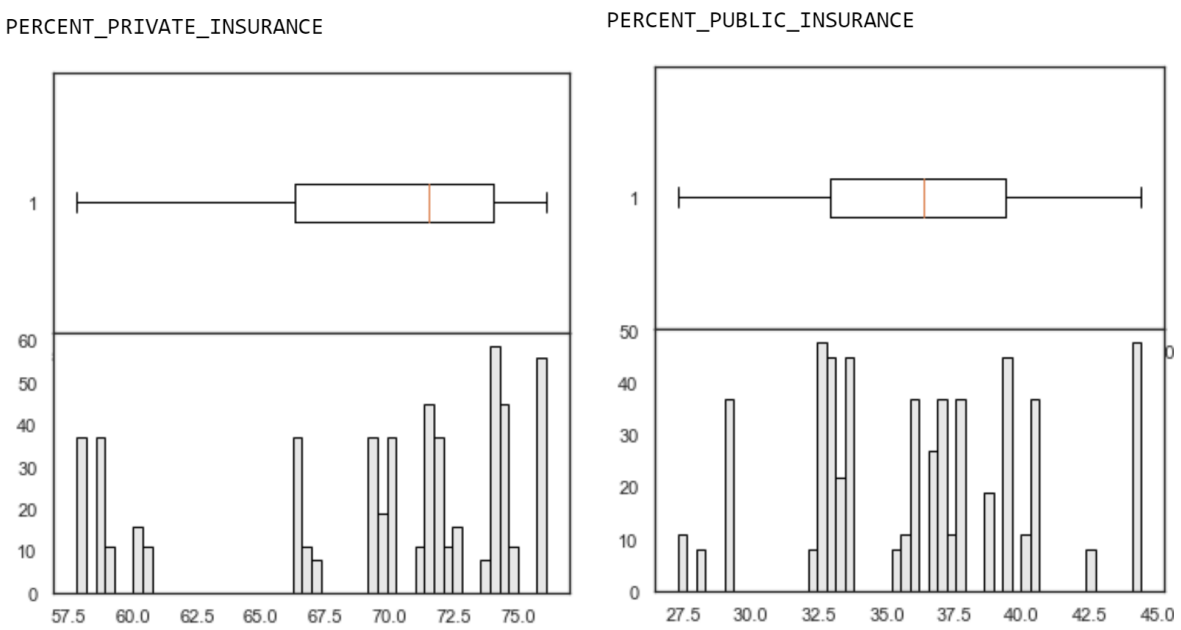


Figure 2. Histogram and Boxplot for Each Attribute

Missing value analysis and outlier analysis

1. VAX_PERCENT_DISTRIBUTION: 151 missing
    a. Cause: There are a few causes of this. Primarily, the CDC has not yet released data for flu season 2021-2022. It will not be released until September of this year; therefore, 100 missing rows of data come from this. Another cause is the CDC does not take measurements in June for any year, given that June is not flu season. A third cause is that we needed to include data from March 2017 for the ESTIMATE % column to calculate April 2017 for the VAX_PERCENT_DISTRIBUTION column. Therefore, we will not have any data from March 2017 in the VAX_PERCENT_DISTRUBTION column. Additionally, there were 12 values in ESTIMATE % that had missing values labeled (NR) by the data.
    b. Solution: For missing data for flu season 2021, we will impute and predict this data when we build our model next quarter. We will delete all June rows for data missing from June since June is not part of flu season anyway. For data missing for March 2017 for each state, we will drop these rows since the data from these rows was only used to calculate April 2017 VAX_PERCENT_DISTRIBUTION. For the 12 rows with "NR" values, if the value was missing for the month of July, this is the start of flu season, so we set this value equal to zero.
    c. Based on the new information that we are missing data from 2021-2022 flu season (this was not known before this report), we will leave the missing values and turn them into a step in our model building. The concept will go as follows:
        i. Step 1: predict 2021-2022 Vax data using google trends + demographic data
        ii. Step 2: forecast 2022-2023 flu season google trends

iii. Step 3: predict 2022-2023 vax data using google trend + demographic

A. First Approach to Other Missing Values

2. PERCENT_OVER_65: 560 missing
3. MEDIAN_HOUSEHOLD_INCOME: 580 missing
4. PERCENT_ATTAIN_BACHELORS: 580 missing
5. Racial Demographic data:
    a. WHITE: 570 missing
    b. BLACK: 570 missing
    c. RACE_OTHER: 570 missing
    d. HISPANIC: 570 missing

- Cause: The US census only takes data measurements and publishes them in July annually/semi-annually. Monthly demographic data were not provided. Additionally, most demographic data for 2021 were also not provided on the US census website by the time this report was written. Therefore, although each demographic attribute is expectedly collected starting from January 2017 to December 2021, we only have data for July of each year for five years. We need to estimate the monthly demographic data for each state based on the current annual data.
- Solution: Several methods were considered to fill in the missing data, such as the average, median, mode, linear regression model, and interpolation. Since the historical demographic data generally increases over time, simply using average, mode, and median to fill in the missing values is not sufficient in our project. Thus, a combination of linear interpolation and linear regression model was used to fill in and predict the missing data.
    - Linear interpolation: Linear interpolation is a form of interpolation that can be used for one-dimensional data. In linear interpolation, a new value is estimated based on the two data points that are in the one-dimensional data sequence with the estimated one (Huang). From our dataset, from Jan 2017 to December 2021, the first non-NA data for demographic data is July 2017, and the last non-NA data is July 2020 (July 2021 for Racial demographic data). By using linear interpolation, the missing monthly data from July 2017 to July 2020 (to July 2021 for Racial demographic data) using the non-NA data were estimated.
    - Linear regression model: linear regression analysis is a linear model used to predict the value of a variable based on the linear relationship with other variables. In our project, certain independent attributes were chosen to predict the missing values. The independent attributes were chosen based on the high correlation between these attributes and features containing missing values. After linear interpolation, the only missing monthly data are from January 2017 to June 2017 and from August 2020/2021 to December 2021. A linear regression model was used to estimate these remaining missing values. Table 2 shows the attributes containing missing values and the corresponding attributes chosen as independent variables for the linear regression model to predict the missing values.

Table 3. Table of Attributes that Were Used in the Linear Regression Model to Predict Missing Values

Attributes Containing Missing Values	Attributes Used to Predict the Missing Value via Linear Regression Model
PERCENT_OVER_65	PERCENT_VOTED_DEMOCRATIC, PERCENT_PUBLIC_INSURANCE
MEDIAN_HOUSEHOLD_INCOME	YEAR, PERCENT_PUBLIC_INSURANCE
PERCENT_ATTAIN_BACHELORS	PERCENT_PRIVATE_INSURANCE, PERCENT_PUBLIC_INSURANCE
WHITE	MEDIAN_HOUSEHOLD_INCOME, PERCENT_PUBLIC_INSURANCE
BLACK	PERCENT_PRIVATE_INSURANCE
RACE_OTHER	MEDIAN_HOUSEHOLD_INCOME, PERCENT_PUBLIC_INSURANCE
HISPANIC	RACE_OTHER, PERCENT_PUBLIC_INSURANCE

Please see Appendix A for the implementation details.

B. Second Approach to Missing Values

The second approach to missing values uses linear regression to predict the future number against the variable Year, and round the result to the nearest percentage. For the Median Household Income, the results are rounded to the nearest thousands.

C. Outliers

Our data before imputation does not have many outliers other than google trends and vaccine distribution data. We want to include these outliers within the google trends and vaccine data because we want to capture the variation in our data- these are the seasonal trends we want to observe. Note that we should not expect outliers by the very nature of the rest of our data. Demographic data does not change too much over time within each state, neither does household income, percent who received bachelors, or voting behavior. There are certainly differences between states (think voting behavior between Wyoming and Pennsylvania); However, the data values within states over time change very slightly, meaning that we do not expect outliers for these variables.

Google Trends. Based on the boxplots, we see outliers for the google trends data. However, this data is already normalized by google, and the outliers are to be expected, as this is the variation of when individuals suddenly start searching for flu shots near them. Therefore, we will keep these outliers.

Vaccine Distribution. We have many outliers for VAX_PERCENT_DISTRIBUTION. Again, we want to capture this high variation in our data, as this corresponds to google trends.

Feature engineering and analysis

Based on the heatmap below, we can see some variables with high correlation. We need to address it in the second phase of the project.

Before coming up with solutions for dealing with high correlation variables, it is essential to interpret why high correlations exist. Some high correlations between variables are explainable because certain variables are mutually exclusive. For example, the percentage voting republican and the percentage voting democrats have a strong correlation because they are a binary choice- the two columns together add up to one. The same principle applies to the demographic data; if there is a greater population percentage of Black people in the state, the population percentage of White people will be lower (Figure 3; Figure 4). We like to conceptualize this as a feature instead of a bug. It is important to understand why certain variables are highly correlated in our data.

In the second phase of the project, we should address high correlations through methods like dimensionality reductions, then benchmark against the data without addressing certain highly correlated variables to see if the high correlation would impact the model's predictive accuracy. So at the current phase of the project, we acknowledge the highly correlated variables and plan to feature engineer them in the second phase of the project.

Note the high correlations amount the google trends data. We acknowledge that the google trends search term that is most highly correlated with VAX_PERCENT_DISTRIBUTION is the search term "FLU SHOT." This is also the term with the largest number of raw google searches relative to the other four terms examined. Note that besides google trends variables, the other highest correlated variables to VAX_PERCENT_DISTRIBUTION are White and Black.

We acknowledge that the columns regarding voting behavior have too high of a correlation (.97). Therefore, they measure the same thing, and one will be removed from the analysis. Also, note that RACE_OTHER and MEDIAN_HOUSEHOLD_INCOME have a correlation of -.96; similarly they seem to be capturing the same thing, though this is a relationship that is a bit less intuitive to figure out.

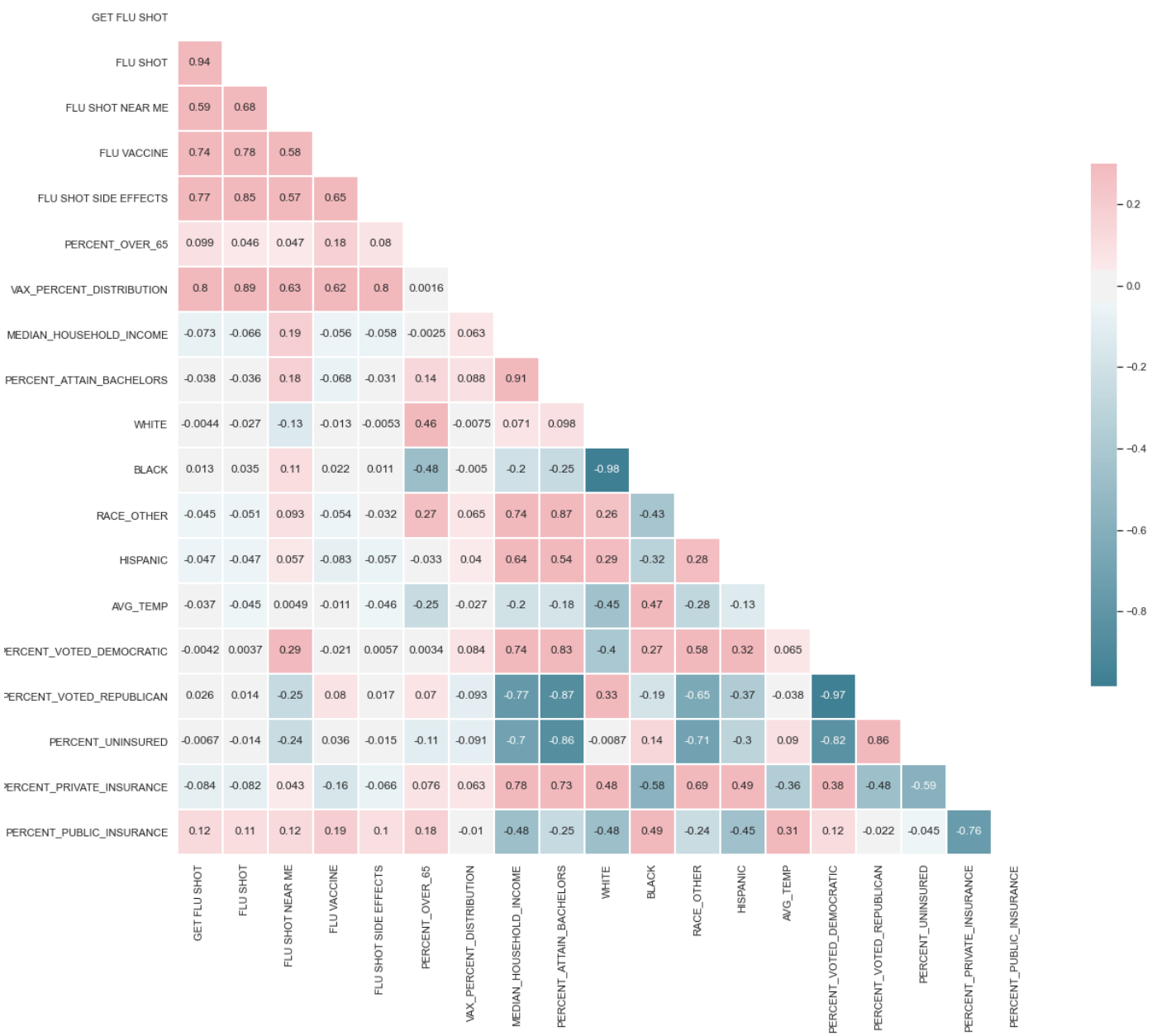


Figure 3. Heatmap for Correlations between Attributes

Although we can see the correlation between any two numerical features in an entire correlation heatmap, we can also filter by a threshold of higher correlation numbers, for example, 80% on the positive and negative sides.

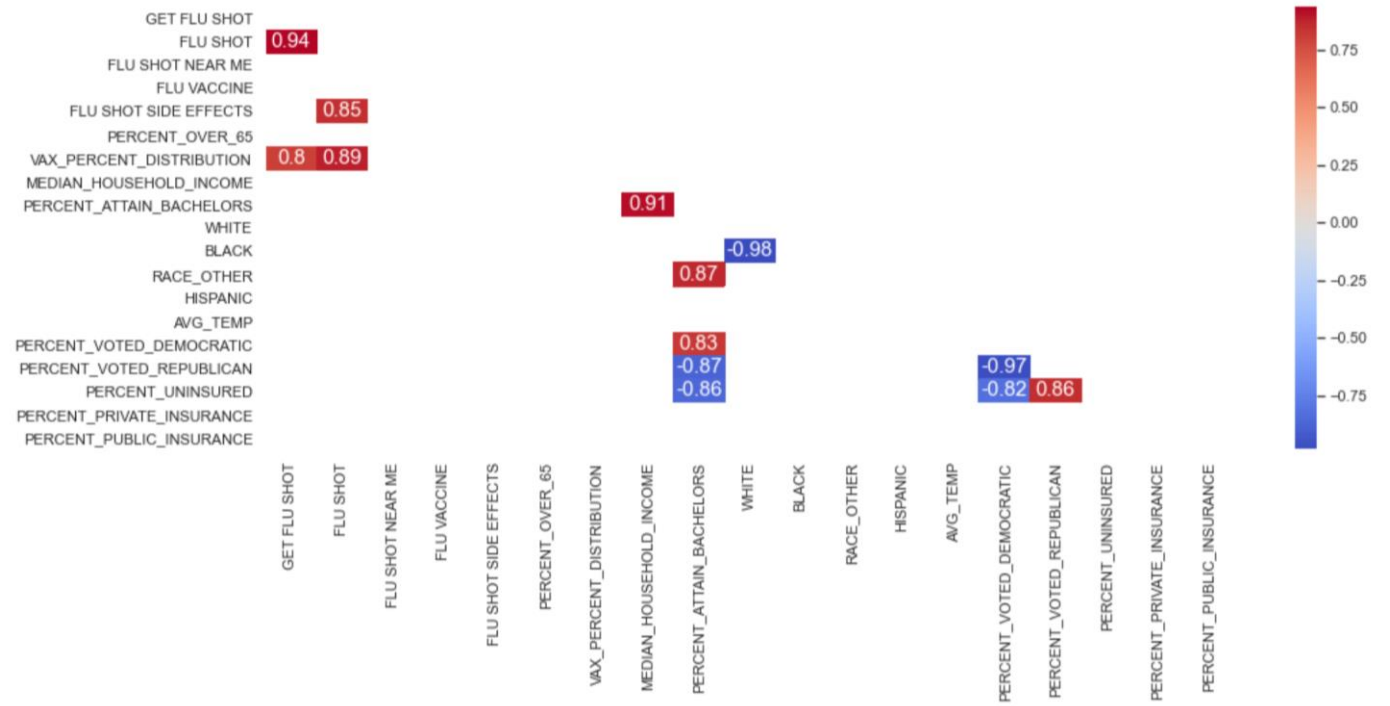


Figure 4. Filtered Heatmap for Correlations between Attributes

Reference:

Huang, G. (2021). Missing data filling method based on linear interpolation and lightgbm.

Journal of Physics: Conference Series, *1754*(1), 012187.

https://doi.org/10.1088/1742-6596/1754/1/012187

Appendix

A. First approach to fill in the missing values: jupyter notebook

B. Summary of dataset **before** missing values were handled

Table 3. Summary of the dataset before missing values were handled

Data Before Missing Values are Handled								
variable	count	mean	std	min	25%	50%	75%	max
GET FLU SHOT	620	3.59	4.04	0.00	1.00	2.00	5.00	24.00
FLU SHOT	620	18.15	21.59	0.00	3.00	8.00	26.25	100.00
FLU SHOT NEAR ME	620	0.66	0.98	0.00	0.00	0.50	1.00	6.00
FLU VACCINE	620	6.36	5.58	0.00	2.00	5.00	9.00	37.00
FLU SHOT SIDE EFFECTS	620	1.28	1.93	0.00	0.00	0.50	1.25	18.00
PERCENT_OVER_65	60	0.17	0.01	0.15	0.16	0.17	0.18	0.20
VAX_PERCENT_DISTRIBUTION	469	5.11	8.04	0.00	0.70	1.80	6.20	51.70
MEDIAN_HOUSEHOLD_INCOME	36	66186	13754	43595	54713	65525	77768	88589
PERCENT_ATTAIN_BACHELORS	40	32.83	7.05	21.90	26.90	32.30	39.63	45.00
WHITE	50	0.78	0.13	0.59	0.63	0.81	0.89	0.93
BLACK	50	0.14	0.14	0.01	0.01	0.10	0.31	0.38
RACE_OTHER	50	0.07	0.03	0.03	0.06	0.07	0.10	0.11
HISPANIC	50	0.09	0.04	0.03	0.05	0.09	0.12	0.17
AVG_TEMP	620	50.79	18.21	3.60	35.68	52.35	66.00	83.50
PERCENT_VOTED_DEMOCRATIC	620	0.46	0.13	0.22	0.38	0.44	0.59	0.66
PERCENT_VOTED_REPUBLICAN	620	0.50	0.12	0.32	0.39	0.53	0.58	0.70
PERCENT_UNINSURED	620	7.74	3.11	2.80	5.50	7.05	10.80	13.00
PERCENT_PRIVATE_INSURANCE	620	69.28	5.82	57.80	66.30	71.50	74.00	76.10
PERCENT_PUBLIC_INSURANCE	620	36.14	4.24	27.30	32.90	36.35	39.30	44.40

C. Summary of dataset after missing value were handled using First Approach

Table 4. Summary of dataset after missing values were handled using First Approach

Summary of dataset after missing values were handled by First Approach								
variable	count	mean	std	min	25%	50%	75%	max
GET FLU SHOT	560	3.92	4.12	0.00	1.00	2.00	6.00	24.00
FLU SHOT	560	19.87	22.02	0.00	4.00	12.00	28.25	100.00
FLU SHOT NEAR ME	560	0.72	1.02	0.00	0.00	0.50	1.00	6.00
FLU VACCINE	560	6.86	5.63	0.00	2.00	6.00	9.00	37.00
FLU SHOT SIDE EFFECTS	560	1.40	1.99	0.00	0.00	1.00	2.00	18.00
PERCENT_OVER_65	560	0.17	0.01	0.15	0.16	0.17	0.18	0.20

VAX_PERCENT_DISTRIBUTION	460	4.33	5.82	0.00	0.60	1.80	5.53	25.30
MEDIAN_HOUSEHOLD_INCOME	560	66897.92	13630.31	43550.02	56065.58	66441.67	79020.38	94221.92
PERCENT_ATTAIN_BACHELORS	560	32.98	6.97	21.90	26.90	32.30	39.75	45.00
WHITE	560	0.78	0.13	0.58	0.63	0.81	0.89	0.93
BLACK	560	0.14	0.13	0.01	0.01	0.10	0.31	0.38
RACE_OTHER	560	0.07	0.02	0.03	0.06	0.07	0.10	0.11
HISPANIC	560	0.09	0.04	0.03	0.05	0.09	0.12	0.17
AVG_TEMP	560	49.44	18.18	3.60	34.70	50.30	65.00	83.50
PERCENT_VOTED_DEMOCRATIC	560	0.46	0.13	0.22	0.38	0.44	0.59	0.66
PERCENT_VOTED_REPUBLICAN	560	0.50	0.12	0.32	0.39	0.53	0.58	0.70
PERCENT_UNINSURED	560	7.74	3.11	2.80	5.50	7.05	10.80	13.00
PERCENT_PRIVATE_INSURANCE	560	69.27	5.83	57.80	66.30	71.50	74.00	76.10
PERCENT_PUBLIC_INSURANCE	560	36.15	4.25	27.30	32.90	36.35	39.40	44.40

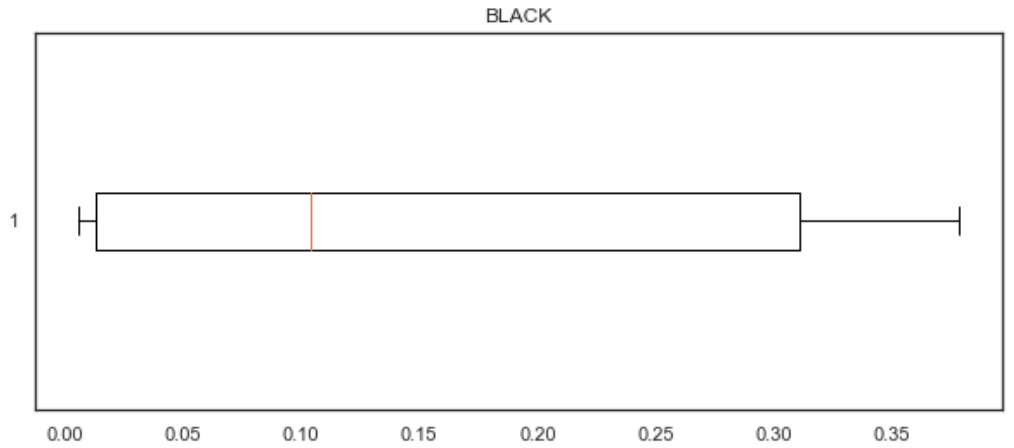
D. Summary of dataset after missing values were handled by Second Approach

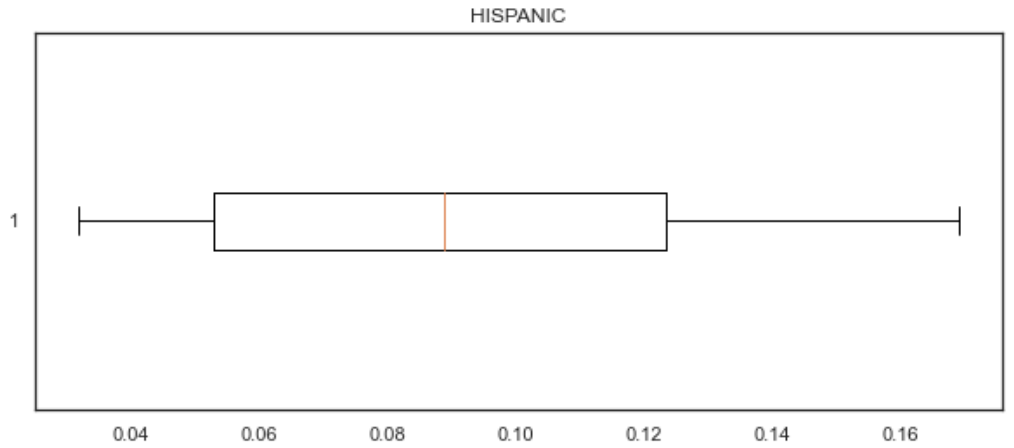
Table 5. Summary of dataset after missing values were handled using Second Approach

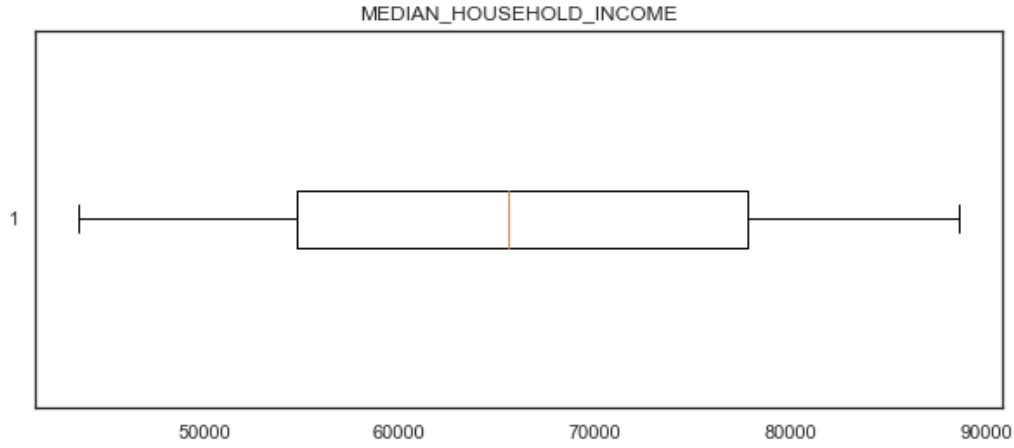
Summary of dataset after missing values were handled by Second Approach								
variable	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
MEDIAN_HOUSEHOLD_INCOME	560	65.595893	14.399558	43.6	51.1	66.2	78.6	91.6
PERCENT_ATTAIN_BACHELORS	560	0.330286	0.06988	0.22	0.27	0.32	0.4	0.45
WHITE	560	0.783	0.125894	0.59	0.63	0.815	0.89	0.93
BLACK	560	0.145	0.134749	0.01	0.01	0.105	0.31	0.38
RACE_OTHER	560	0.072607	0.025236	0.03	0.06	0.07	0.1	0.11
HISPANIC	560	0.088018	0.042281	0.03	0.05	0.09	0.12	0.17

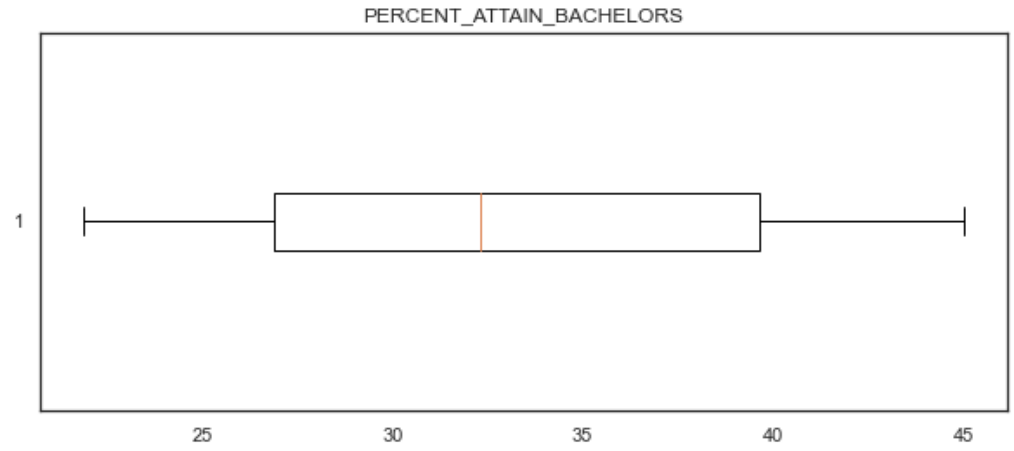
E. Histograms of data prior to interpolation

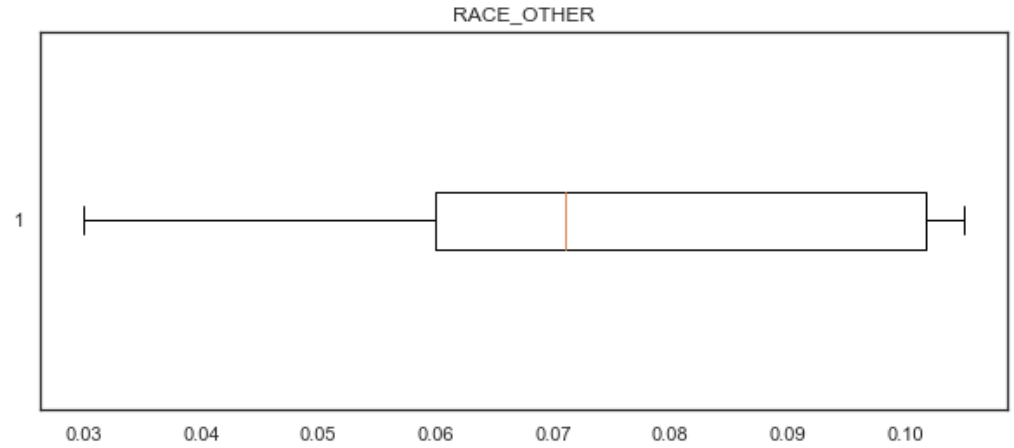
F. Box Plots of 6 variables (Race, Bachelors, Median Income) prior to missing data

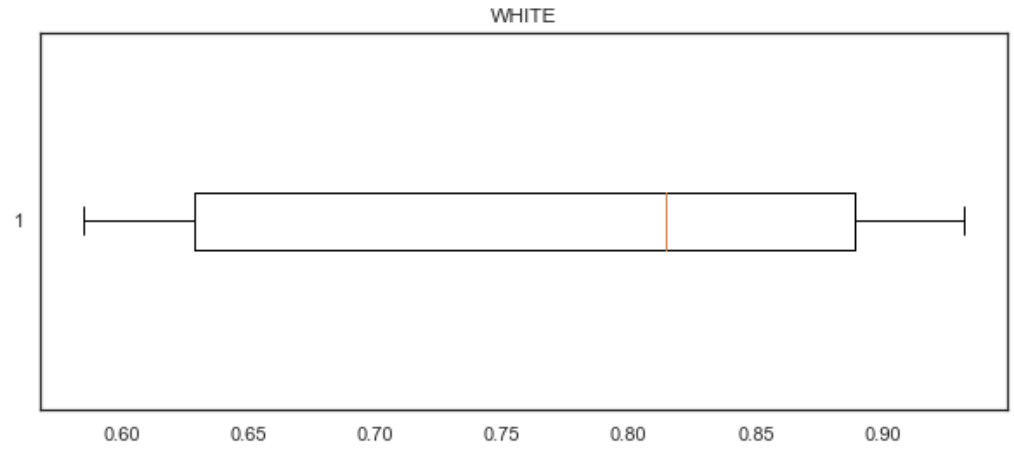






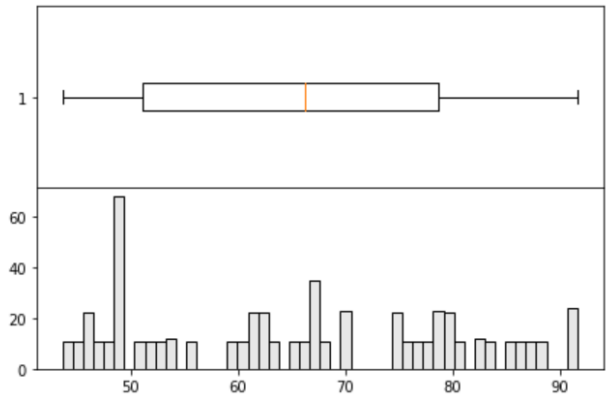

PERCENT_ATTAIN_BACHELORS
1
25
30
35
40
45


RACE_OTHER
1
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10

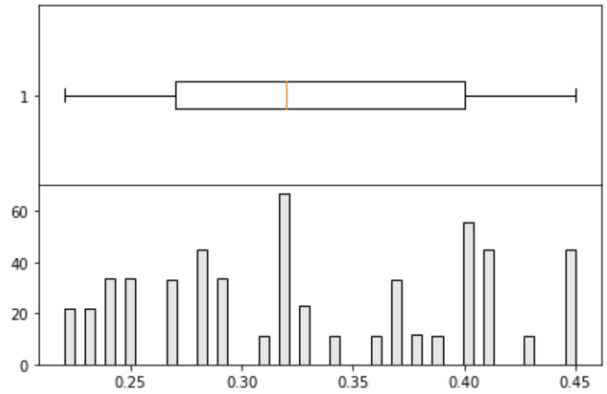

WHITE
1
0.60
0.65
0.70
0.75
0.80
0.85
0.90

G. Box plots of 6 variables (Race, Bachelors, Median Income) after Approach 2

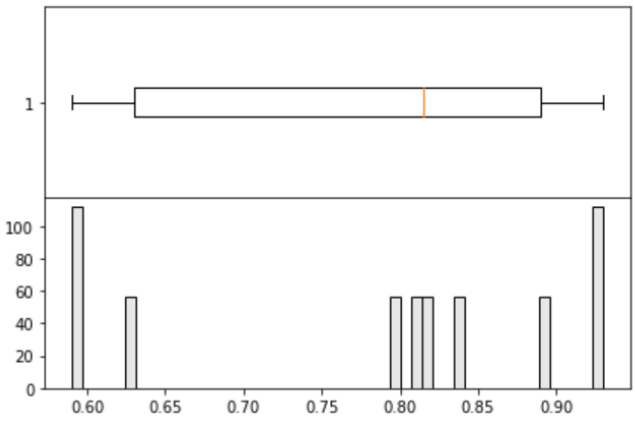
MEDIAN_HOUSEHOLD_INCOME:



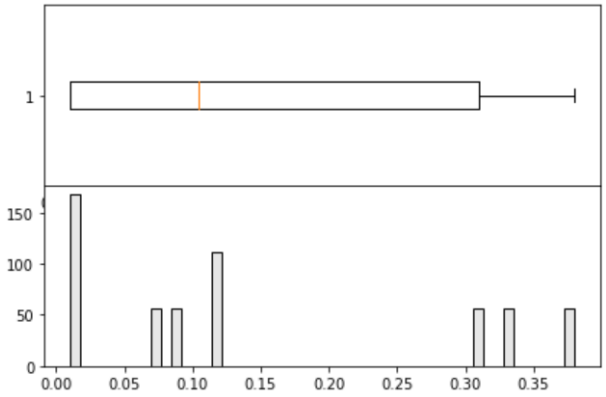
PERCENT_ATTAIN_BACHELORS:



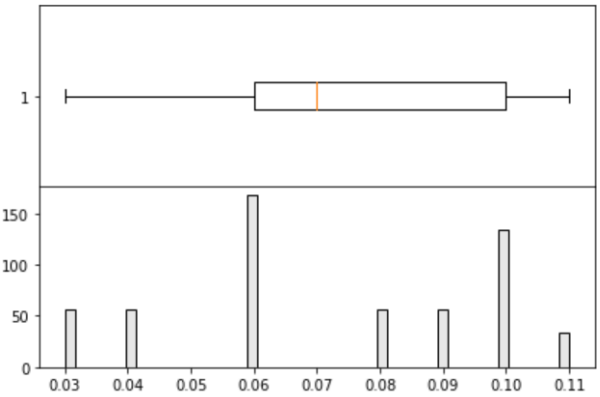
RACE: WHITE:



RACE: BLACK:



RACE: OTHER:



RACE: HISPANIC: