Final Project - UAS Sightings and Registrations

Winter Institute in Data Science

Mark, Marc, Jocelyn, Sophie

2020-1-11

Abstract

Our goal was to analyze the correlation of Unmanned Aircraft Systems (UAS, or drone) sightings in each state to the amount of registrations of UAS + other variables from 2016 to 2018. We were interested in the drone data and the policies that may come with the correlation between registrations and sightings. we also looked into other factors (population, income, state rankings on entrepreneurship and/or technology and science). The assumption that UAS sightings and registrations correlate positively was held when only accounting for those two variables, but when controlling for population and incorprating additional factors into our model, we see that those other factors become more signficant.

Data Sources

We used a variety of sources, all websites, to accrue data sets of our characteristics that prompted us to test the correlation of Unmanned Aircraft Systems (UAS) sightings in each state to the amount of registrations in the same state from 2016 to 2018; UASs are commonly known as drones. Our main data source was the Federal Aviation Administration (FAA) website that gave us the data sets of the total number of reported sightings Federal Aviation Administration (n.d.b), broken down to the very day and time of the report, and the total number of new registrations Federal Aviation Administration (n.d.a) of UASs in each state, by the year. We combined the daily data of reported sightings and accounted for the yearly total in our data set in R, so we could have it at the same standard as registrations. We also combined "hobbyist" and "non-hobbyist" registrations for the purpose of our analysis. Hobbyist registrations are covered under Section 366 of the FAA website, and is for recreational flyers and modeler community-based organizations; the individual registered is considered a "modeler" and must be 13 years of age and a US citizen or legal permanent resident. Under Part 107 of Section 366, certified remote pilots, including commercial operators, must register. Our other two data sources were the Census Bureau and the US Bureau of Economic Analysis (BEA). The Census data sets Bureau (n.d.) measured the income and population of the United States, at the state level while the BEA accounted for population and income per county, by state. Economic Analysis (n.d.) Once we started to look at different characteristics, we found that there was a significance between the number of sightings and the number of entrepreneurial activity in the United States per state; we used the Kauffman Indicators of Entrepreneurial website and their 2018 index level. ("Kauffman Early-Stage Entrepreneurship (KESE) Index" 2018) Beyond the entrepreneurship index, we looked at the Milken Institute collection 2018 science and technology state rankings to see if there was a more in depth connection to our original variables, registrations and sightings. ("State Tech and Science Index" n.d.) Further, we tried to incorporate weather statistics, assuming that if a state was known to have more "clear" days per year, they would have more sightings but we could not find an accurate data source that accounted for data that aligned with our other characteristics. We found a data set on the Current Results website, but it was only data for major cities in each state, assuming the amount of sunshine a state usually has in a year.

Definitions

"Sightings" are reports made my citizens, pilots, and law enforcement to the FAA regarding any unmanned aircrafts sightings across the nation from 2016 to 2018. The FAA has reported that sightings have increased dramatically over the last two years, receiving more than 100 such reports each month. Although the FAA has continued work with their partners through the "Know Before You Fly" campaign to educate the public on how to operate within the rules and laws, unauthorized flying has remained an issue. Working closely with law enforcement, the FAA tries to identify and investigate UASs operations and rely heavily on the sightings report to do so.

"Registrations" are the FAA's way of identifying and tracking the owners, operators, and flight patterns of UASs across the nation. Since the UAS will be flown in the National Airspace System, the FAA provides a list of safety tips for all operators, as well as the standards for different types of operators: defined as either hobbyist and non-hobbyist. Registrations cost $5 per aircraft and is valid for 3 years. For both hobbyist and non-hobbyist registrations, which we combined in the data, a register must provide an email address, credit or debit card, physical and mailing address, and the make and model of the unmanned aircraft.

Sources of Bias

In terms of questions asked, the first bias is probably from each of our background. Our group is interested in finding the correlations between drones registration and the number of drone sightings reported to FAA, however, not everyone is interested in this topic. So, our selection of questions to explore is an obvious bias. The next potential bias is the measure we used. The original data we found range from 1990s to 2018. Due to the fallacies of the initial data, we chose data that aligned with our parameters. As to the data gathering, we chose the parameters of the year (2016-2018). We tried various variables to experiment on what kind of results we can get and we decided to measure the sightings by state. In addition, data gathered by each website doesn't have a clear way of how they collected data. Other factors that may lead to biases including outliers, in our analysis, it is the California. Some missing values include some sightings but people may not report, and also drones that are registered. Finally, because we did not account for a random variable, therefore, we did not make corresponding interpretations. Another bias in data analysis is the regarding the reproducibility. We do not consider we have bias in reproducibility because there are no random variables that will change when the code is running each time. We added the Kauffman Indicators of Entrepreneurship and Milken Institute Science and Technology Rankings for another input. The variables that created the entrepreneurial index and income per state that isn't accounted for in our data set.

Data Cleaning

The data for our project came from four separate sources: One (1) US Census State Population (Yrs 2016-2018), One (1) US Census Median Income (Yrs 2016-2018), sixteen (16) Federal Aviation Administration Drone Sightings, Five (5) Federal Aviation Administration Drone Registrations

Data from these URL links (saved online as Microsoft Excel .xlsx files) into an R Markdown file, as well as the necessary R packages (tidyverse, dplyr, stringr, DescTools, etc). The data was then cleaned in R, in order to merge into a consistent format:

The population data and income data (from US Census site) was already in an easy format to read into R with separate rows for each State (as well as District of Columbia). Some minimal cleaning was required, and included indexing certain row/column names, and changing column names for consistency. Also, the syntax of column names was modified using various functions to be a consistent format for merging.

In regards to dates, they were in various formats (e.g. Jun 1, 2019 or 2019-06-01, or Jun 1, 2019 08:33 or missing) for both of the files. For the drone sightings database, the dates were specific to the day, while the drone registration database was specific to the quarter of the year. Thus, this drove the decision to examine the data primarily from a year-by-year level.

In regards to locations, the drone registration database was in a City and State format, while the drone registration data was in City, Zip Code and State format. In addition, there was missing City and Zip code data for a chunk of the drone registration data. Since it was difficult to locate population and income data specific to an individual City and Zip Code, the data was summarized primarily at a State level for location. In addition, the locations were in various formats (VIRGINIA, Virginia, VA), which required processing to get in consistent format. Data which contained missing date or location values were removed from the dataset. After merging/binding the data, it provided columns for median income, population, drone registration in 2016, drone registration in 2017, drone registration in 2018, drone sightings in 2016, drone registration in 2017, and drone sightings in 2018.

Exploratory Data Analysis

First we measured correlation of number of UAS sightings and registrations by year.

```
cor(sight16, reg16)
```

```
## [1] 0.8553253
```

```
cor(sight17, reg17)
```

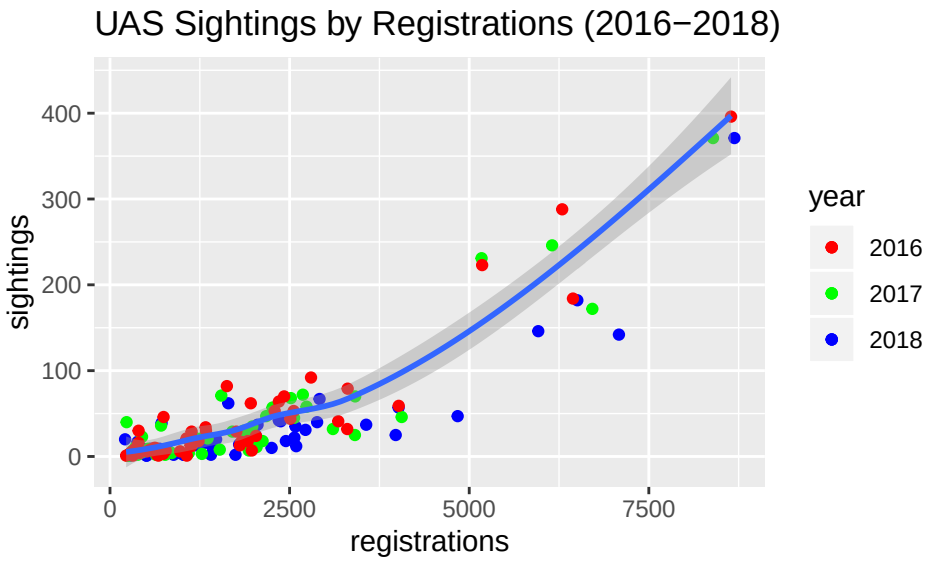
```
## [1] 0.8917381
```

```
cor(sight18, reg18)
```

```
## [1] 0.912211
```

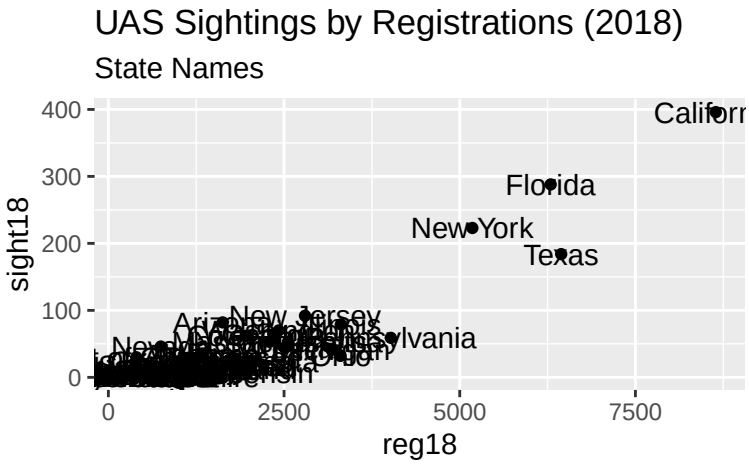
An interesting trend we noticed is that correlation between the two over the past 3 years have steadily increased. We then wanted to see this correlation on a plot:

```
# Plot UAS registrations and sightings for '16, 17, 18
ggplot(data = dataTOTAL) +
  geom_point(mapping = aes(x=reg16, y=sight16, color = "red")) +
  geom_point(mapping = aes(x=reg17, y=sight17, color = "green")) +
  geom_point(mapping = aes(x=reg18, y=sight18, color = "blue")) +
  geom_smooth(aes(x=reg18, y=sight18)) +
  xlab('registrations') +
  ylab('sightings') +
  scale_color_manual(name = 'year',
         values =c('green'='green', 'red'='blue', 'blue'='red'),
         labels = c('2016', '2017', '2018')) +
  labs(title = "UAS Sightings by Registrations (2016-2018)")
```



Exploring this data seems to visually suggest that there is indeed a strong correlation between number of sightings and registrations. We added the state names to the plot below.

```
ggplot(data = dataTOTAL) +
  geom_point(mapping = aes(x=reg18, y=sight18))+
  geom_text(aes(x=reg18, y=sight18, label=State)) +
  labs(title = "UAS Sightings by Registrations (2018)",
       subtitle = "State Names")
```



However, we then thought about how there might be other factors that may be impacting number of sightings by registrations. One was population - if we see the top states by number of sightings or registrations, the same 4 states appear and happen to also be the same outliers in the above plot.

The plot below shows the same plot as above but with sightings and registrations for every 1000 people.

```
headSight <- dataTOTAL %>% select(State, Population, SightCount2018)
headReg <- dataTOTAL %>% select(State, Population, RegCount2018)
head(arrange(headSight,desc(SightCount2018)), n = 4)
```

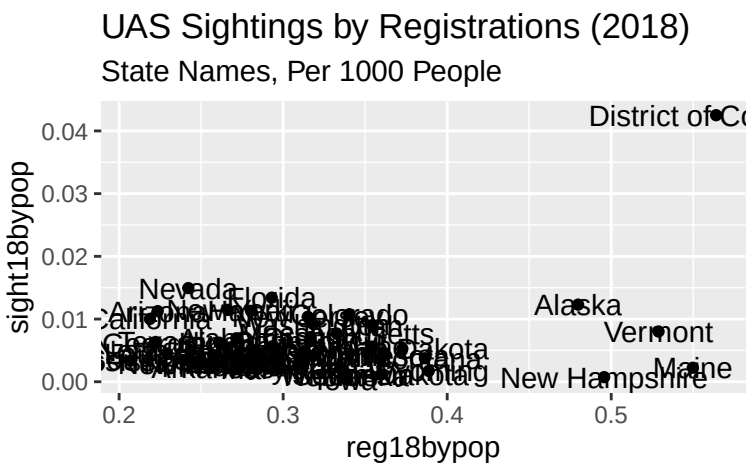
```
##        State Population SightCount2018
```

```
## 1 California   39512223            396
## 2    Florida   21477737            288
## 3   New York   19453561            223
## 4      Texas   28995881            184
```

```
head(arrange(headReg,desc(RegCount2018)), n = 4)
```

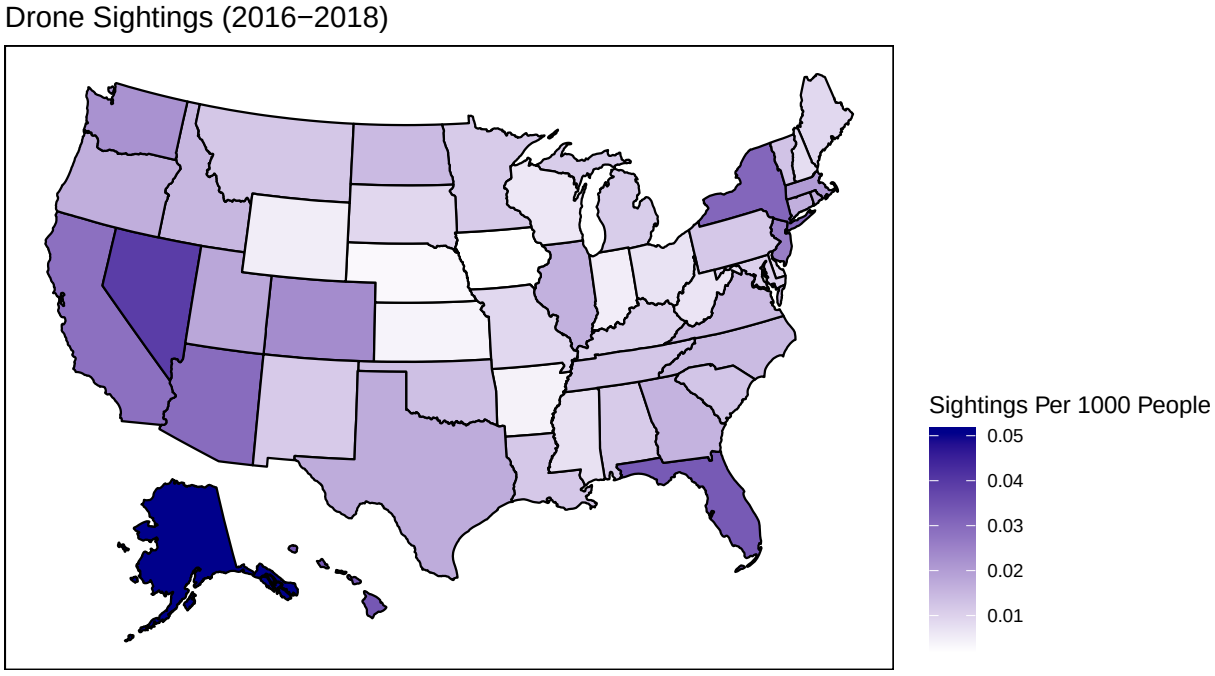
```
##        State Population RegCount2018
## 1 California   39512223         8644
## 2      Texas   28995881         6444
## 3    Florida   21477737         6294
## 4   New York   19453561         5180
```

```
ggplot(data = dataTOTAL) +
  geom_point(mapping = aes(x=reg18bypop, y=sight18bypop)) +
  geom_text(aes(x=reg18bypop, y=sight18bypop, label=State)) +
  labs(title = "UAS Sightings by Registrations (2018)",
       subtitle = "State Names, Per 1000 People")
```



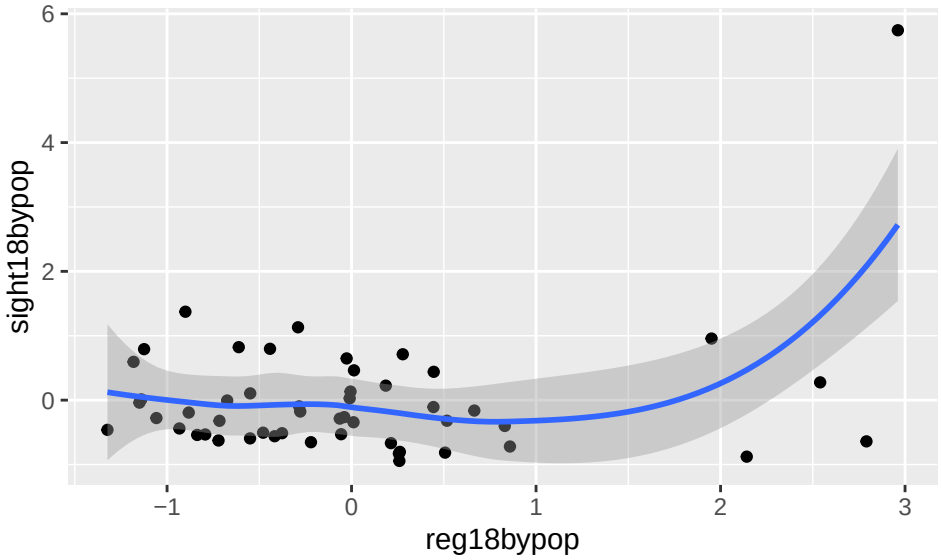
Below is a geographic visualization of the above data on the USA.

```
# Maps total sightings 2016-18 by pop per 1000 on USA
# Removed DC as an outlier for visualization purpose
dataTOTAL_noDC <- dataTOTAL[dataTOTAL$State != "District of Columbia", ]
names(dataTOTAL_noDC)[1] <- "state"
dataTOTAL_noDC$RelSightings <- (dataTOTAL_noDC$TotalSightings/dataTOTAL_noDC$Population)*1000
plot_usmap(data = dataTOTAL_noDC, values = "RelSightings") +
labs(title = "Drone Sightings (2016-2018)") +
theme(panel.background = element_rect(colour = "black")) +
scale_fill_continuous(low = "white", high ="darkblue",
name = "Sightings Per 1000 People") +
theme(legend.position = "right")
```

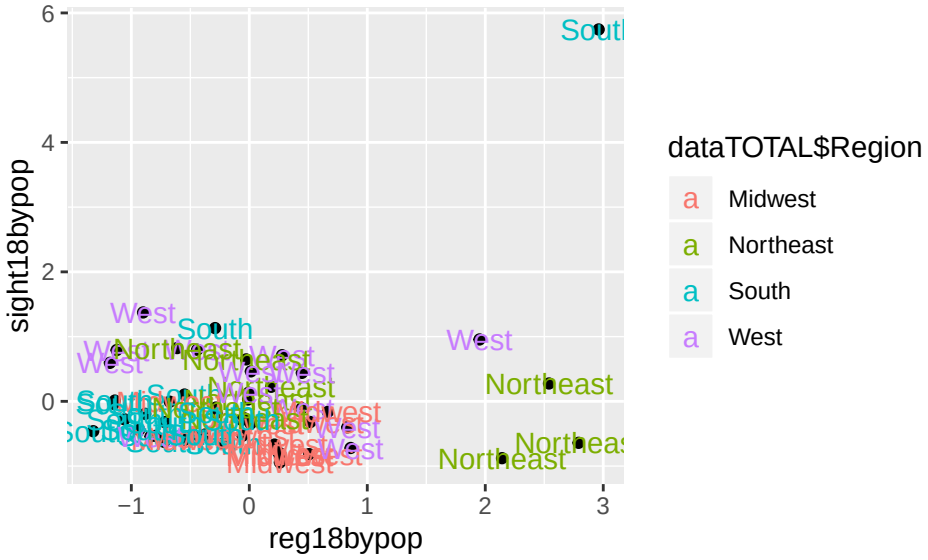


We wanted to explore if there was anything interesting in the clusters of the data and US Census regions. First we standardized the data for a k-means clustering, and first visualized the data based on regions without the k-means clustering. We than visualized the clusters (we used 5 clusters to represent the 4 regions, with DC being its own cluster as an outlier).

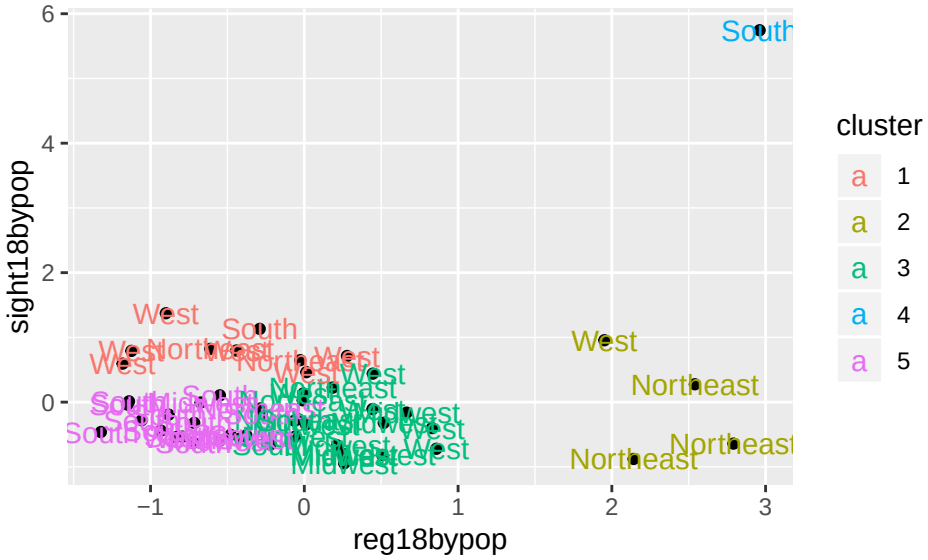
```
kdata <- dataTOTAL %>% select(SightCount2018, RegCount2018) %>%
  transmute(sight18bypop) %>%
  mutate(reg18bypop)

kdata.standard <- data.frame(scale(kdata))
ggplot(kdata.standard, aes(reg18bypop, sight18bypop)) + geom_point() + geom_smooth()
```




```
# k-means clustering
cls <- kmeans(kdata.standard, 5)
kdata.standard$cluster <- as.character(cls$cluster)
ggplot(data = kdata.standard) +
  geom_point(mapping = aes(reg18bypop, sight18bypop)) +
  geom_text(aes(x=reg18bypop, y=sight18bypop, label=dataTOTAL$Region, colour = dataTOTAL$Region))
```




```
ggplot(data = kdata.standard) +
  geom_point(mapping = aes(reg18bypop, sight18bypop)) +
  geom_text(aes(x=reg18bypop, y=sight18bypop, label=dataTOTAL$Region, colour = cluster))
```



This is purely exploratory but we were interested in seeing if the regions would make up most, if not all, of a cluster generated by the k-means. The graph shows that the states within each region were not all similar to one another. We also found that no matter the number of clusters we generated, DC remained an outlier and its own cluster.

Linear Regression

We ran one linear regression models to see if the outcome (sightings by population) are affected by registrations by population, and ran another where we included explanatory variables of median income, Kauffman Index, and Milken Ranks.

```
summary(lm(sight18bypop ~ reg18bypop, data=dataTOTAL))
```

```
##
## Call:
## lm(formula = sight18bypop ~ reg18bypop, data = dataTOTAL)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.009351 -0.002787 -0.001115  0.001892  0.030963
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0005074  0.0034008  -0.149   0.8820
## reg18bypop   0.0213716  0.0103754   2.060   0.0447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006113 on 49 degrees of freedom
## Multiple R-squared:  0.07969,    Adjusted R-squared:  0.06091
## F-statistic: 4.243 on 1 and 49 DF,  p-value: 0.04475
```

When we are only measuring the comparison between sightings by population and registrations by population, we again see the positive correlation and a significant p-value (< 0.05). However we also wanted to explore

if other variables such as Median Income from BEA, the Kauffman Index on Entrepreneurship or Milken Index on State Science and Technology Rankings had impact on the sightings by population.

```
summary(lm(sight18bypop ~ reg18bypop + Med_Income + zindex18 + milrank, data=dataTOTAL))
```

```
##
## Call:
## lm(formula = sight18bypop ~ reg18bypop + Med_Income + zindex18 +
##     milrank, data = dataTOTAL)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0056362 -0.0015625 -0.0004673  0.0018227  0.0082714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.652e-04  3.269e-03  -0.295  0.76914
## reg18bypop  -1.196e-02  6.856e-03  -1.744  0.08797 .
## Med_Income   1.646e-07  7.587e-08   2.170  0.03536 *
## zindex18     1.088e-03  3.733e-04   2.914  0.00554 **
## milrank     -2.916e-04  4.127e-03  -0.071  0.94399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003223 on 45 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2811, Adjusted R-squared:  0.2172
## F-statistic: 4.399 on 4 and 45 DF,  p-value: 0.004363
```

Now we see that the most significant variable is interestingly the Entrepreneurship Index. We assumed (when we decided to add the Milken rankings) that perhaps entrepreneurship may have been too vague to explain drone-related activities in a state, which is why we wanted to look at something more specific like Science and Technology rankings. However this regression shows that the Milken rankings are not significant at all in this model. Median Income was also significant, but not as much as the entrepreneurship index.

Next Steps

Given more time, we recommend performing additional data cleaning and data analysis. For data cleaning, we recommend performing an Amelia method to help predict and fill in missing values for data, particularly location and time data for drone registrations. In addition, the drone sighting data did have some time stamp data, which could be extracted into a separate column to analyze drone sighting time data. Thus, we could have added drone sighting time data as a predicting variable in the regression model and other future analysis. In addition, we could reach out to the FAA or explore the registration data set more closely to determine if there is a way to determine drone registrations by day instead of quarter year (to compare to daily drone sightings). Since the drone registration data was given as a result of Freedom of Information Act (FOIA) request, another request could be made to understand the data in terms of days. For the data analysis, we could perform other machine learning analysis in addition to regression modeling. The clustering method was explored, but could be tweaked to remove outliers and explore segments of States that meet a certain condition. For example, there could be a cluster of Western states with high drone registration data for a given reason. In addition, decision trees, random forest, neural networks and other machine learning methods could be explored against the regression model. However, many of these analysis are difficult with a small dataset (e.g. 51 states in addition to D.C.), and we could explore changing the structure of our data to provide more records for analysis.

References

Bureau, The United States Census. n.d. "State Population Totals: 2010-2019." Accessed January 9, 2020. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html.

Economic Analysis, U. S. Bureau of. n.d. "Personal Income by County, Metro, and Other Areas." Accessed January 9, 2020. https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas.

Federal Aviation Administration. n.d.a. "FOIA Library." Accessed January 9, 2020. https://www.faa.gov/foia/electronic_reading_room/.

———. n.d.b. "UAS Sightings Report." Accessed January 9, 2020. https://www.faa.gov/uas/resources/public_records/uas_sightings_report/.

"Kauffman Early-Stage Entrepreneurship (KESE) Index." 2018. *Kauffman Indicators of Entrepreneurship*. https://indicators.kauffman.org/.

"State Tech and Science Index." n.d. Accessed January 11, 2020. http://www.statetechandscience.org.