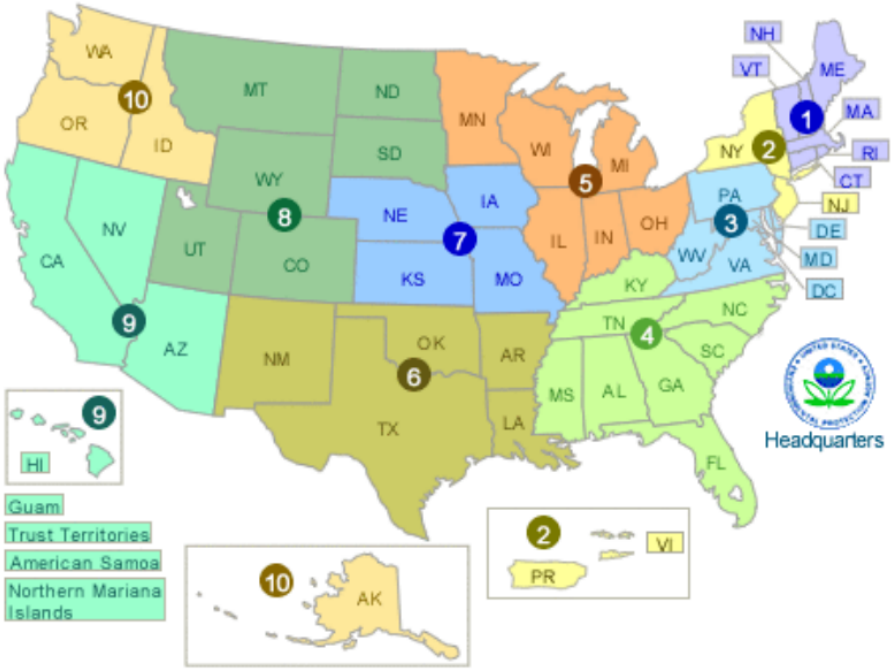
CPLN 505 Assignment 2

Shadrack Lilan

Philadelphia is served by a combined sewer system and is currently considering efforts to improve it's sewer system in a manner that will provide environmental and economic benefit to the city. We will look at the environmental and population data from different cities and facilities in the US and hopefully glean some useful information that we can use to guide Philadelphia's decision.

Throughout this report, we will be relying on official data from the EPA and the US census office. Because there are many facilities, most of the analysis will be done in regional chunks for ease of analysis and representation. The EPA Regions are labelled from 1-10 and are indicated below:



EPA Regions 1-10. Source: usepa.gov

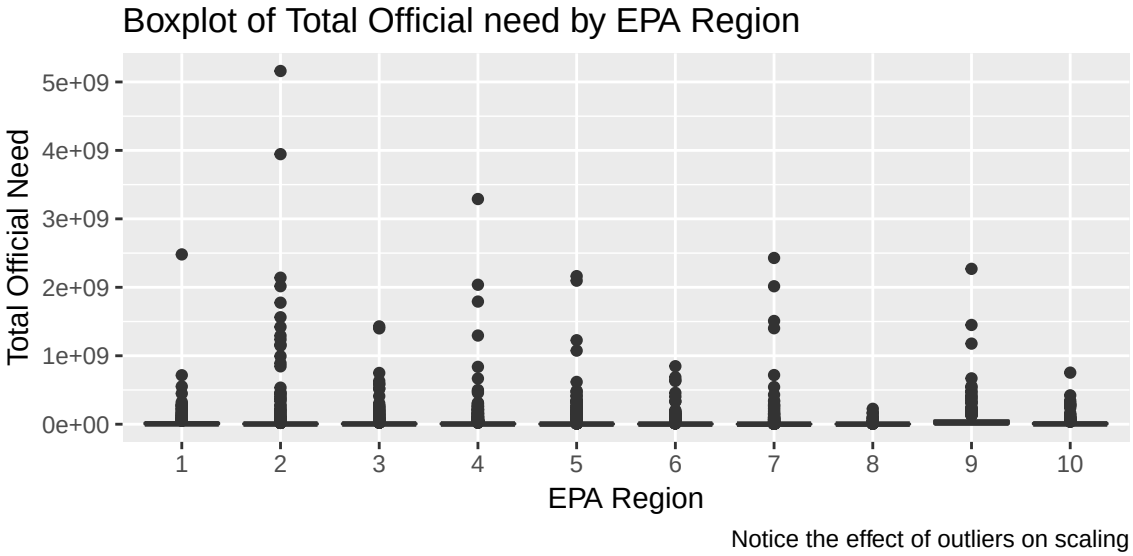
Section 1: Summary Statistics and visual plots

We begin the analysis by exploring certain variables in the in the dataset. We will also use plots and other visual representation to see the general trend

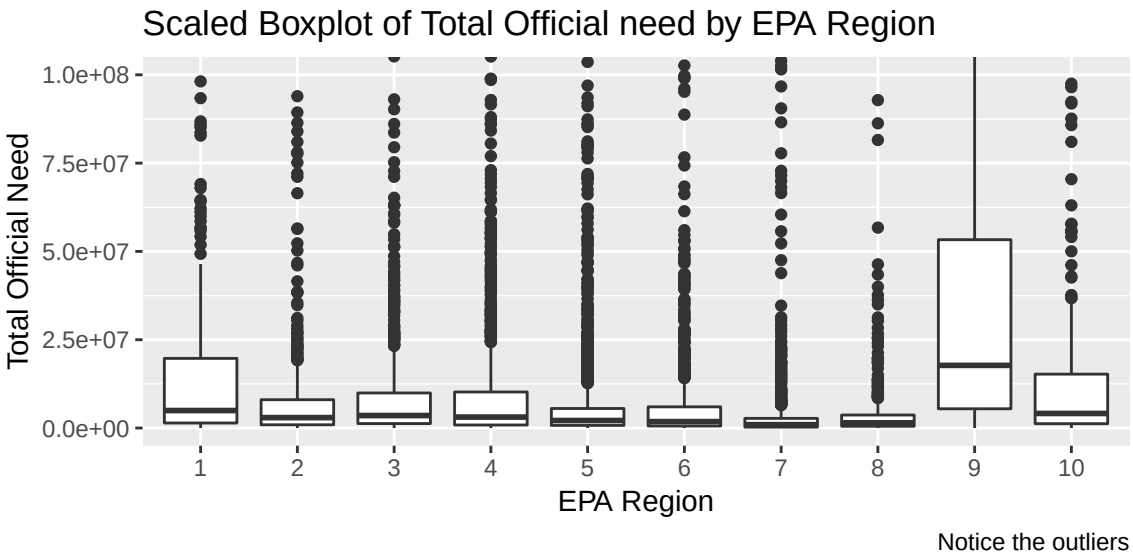
1. Summary statistics of test data.
We will start by looking at some summary statistics of some variables of interest. Philadelphia facilities seem to have a high need, therefore, I will look at some variables that correspond to philadelphia's location and demographics. The variables under consideration are:
- TOTAL_OFFICIAL_NEED
- TMDL_INDICATOR - POP10 - PCTWHITE10 - MEDINC09

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
TOTAL_OFFICIAL_NEED	0.0	704598.8	2409821.50	2.162322e+07	7660571.0	5160108526
POP10	494.0	26634.5	74319.00	3.047271e+05	270056.0	9818605
MEDINC09	14.2	74.4	87.35	8.249095e+01	94.3	99
PCTWHITE10	22053.0	48855.0	56711.00	5.907579e+04	65587.0	129326

The table above shows different summary statistics for some of the variables, however, an interesting observation is the that the mean and median for TOTAL_OFFICIAL_NEED and POP10 have very high means relative to their median value. A possible explanation for TOTAL_OFFICICAL_NEED is that there are some facilities that serve very huge metropolitan areas including suburbs. A fair share of these are located in FL, NY and CA, and this could be because they are fairly densely populated regions. This is also confirmed by the initial boxplots done for the data that show how some of the points being outliers/high leverage points.



From the boxplot above, we can tell that the region with the facility that had the highest need would region 2. This region inlcudes the State of New York, and it would make sense that New York City has the facility with the highest demand. However, these boxplots do not convey alot of information, therefore we can scale these boxplots to give a better visual of what's occuring region. A sidenote is that Philadelphia is in region 3, which exhibits a moderate need relative to the rest of the regions.



An improved boxplot shows us a slightly better representation of some summary statistics per EPA region. A quick look confirms the earlier point that Philadelphia's region 3 falls right in the middle of the pack. The scaled in boxplots also show that region 9 has a higher dispersion than other regions. This could be because California bears the brunt of the population in this region, hence may account for most of the need compared to Arizona or Nevada, which are known to be sparsely populated, hence less pressure on sewer facilities.

A closer look at the activity in each facility might give us more insight on sewer systems in highly need areas, and those that are in low need areas.

Section 2: Tests of association and correlation

In this section we would want to see if there is some association between some of the categorical variables and the continous outputs. This would be helpful to figure out if there are some categories within the data set that might inform us

The first association test will be between facilities' total need with a combined sewer system (CSS) or with a Municipal Separate Sewer System(MS4s).

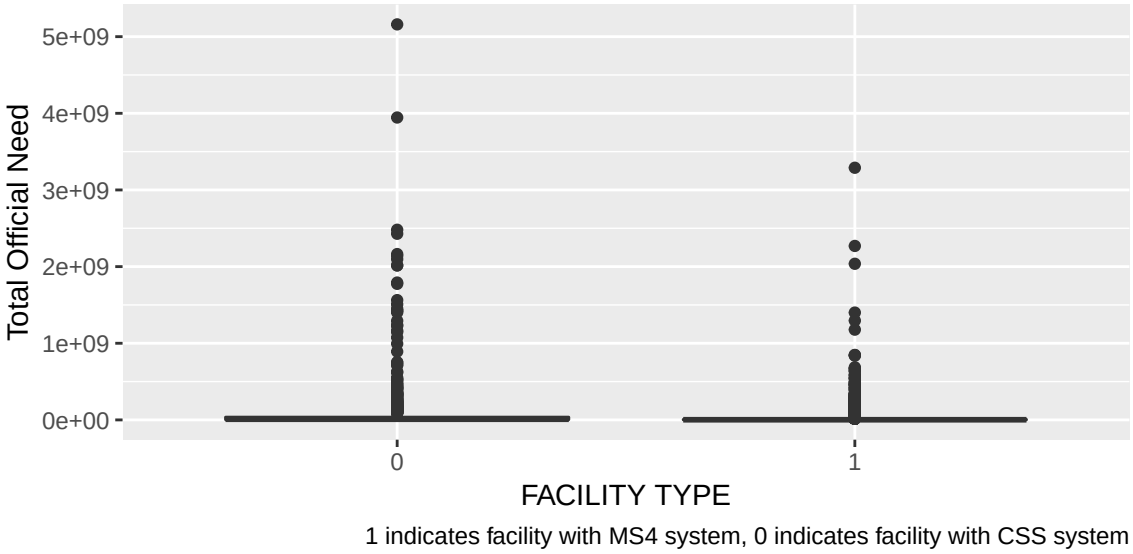
The summary data for facilities with MS4 and CSS is:

	minimum	q1	median	mean	q3	maximum
MS4	0	628112.5	2161876	13384564	6708263	3289580276
CSS	0	2832885.2	8813826	113053301	36254645	5160108526

The summary data for facilities with CSS is:

A look at the mean and the median of the two categories indicates that facilities with combined sewer systems have a higher median, mean and maximum total official need. This could indicate that there is a bit of concern with systems that have combined sewer systems. The summary statistics are also surprising considering that the number of CSS facilities is several factors smaller than the number of MS4 facilities (696 **CSS** facilities compared to 7724 **MS4** facilities).

We will use a boxplot to see if there number of outliers may affect the descriptive statistics for the two types of facilites:



The boxplots indicate presence of outliers in both systems, however it also appears that the facilities with the most need for infrastructure are the ones with a combined sewer system.

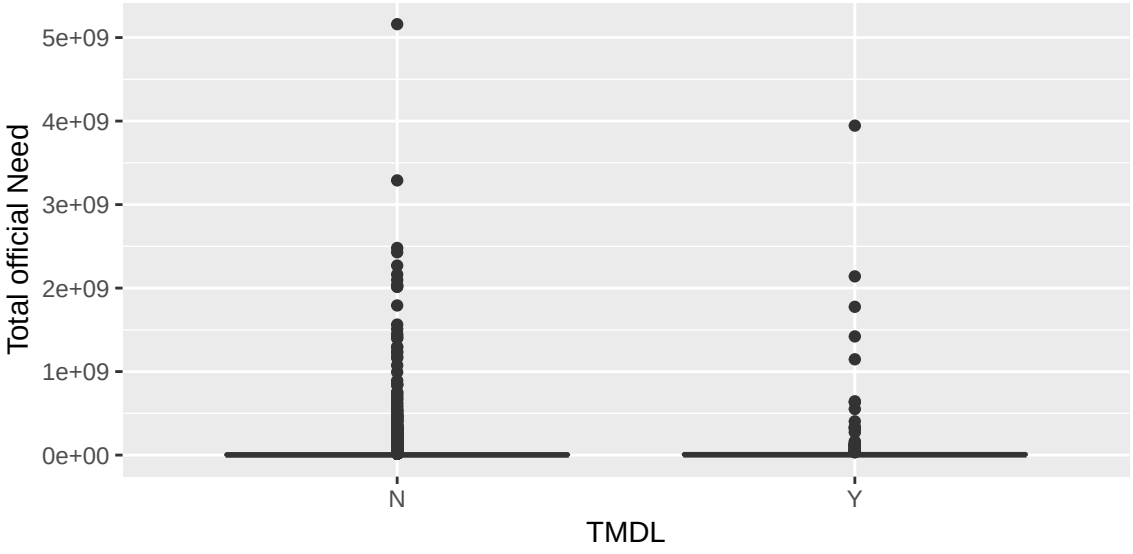
We will do a a t-test to see if there is a significance in the difference of the means of the two sewer systems.

```
## # A tibble: 1 x 3
##    estimate statistic  p.value
##       <dbl>     <dbl>    <dbl>
## 1 99668737.      6.71 3.97e-11
```

With a p-value that is well below the 0.01 threshold, we can reject the null hypothesis and infer that the difference in means of the two groups is statistically significant. This means that we can say that the choice of categorical variable is statistically signficant in the dataset.

Part 2:

We will then proceed to investigate if there is a significant difference between facilities' need for those that are associated with a TMDL. Afterwards we will check if these vary by region.



	summary__yes__tmdl
Min.	0.000e+00
1st Qu.	1.249e+06
Median	3.862e+06
Mean	5.900e+07
3rd Qu.	1.437e+07
Max.	3.945e+09

	summary__no__tmdl
Min.	0.000e+00
1st Qu.	6.908e+05
Median	2.372e+06
Mean	2.023e+07
3rd Qu.	7.537e+06
Max.	5.160e+09

The next step is to test if there is a significanct difference in the means. We can accomplish this using a t-test on the two quantities.

```
## # A tibble: 1 x 3
##     estimate statistic p.value
##        <dbl>     <dbl>   <dbl>
## 1 -38768634.     -2.23  0.0266
```

We obtain a p-value of 0.02, therefore we can reject the null hypothesis and infer that the difference in the two means is significant. The next step is to see the manifestation of this difference by region.

```
## # A tibble: 1 x 3
##     estimate statistic p.value
##        <dbl>     <dbl>   <dbl>
## 1 115714875.     0.859   0.453
## # A tibble: 1 x 3
##      estimate statistic p.value
##         <dbl>     <dbl>   <dbl>
## 1 1754540487.      3.52  0.0168
## # A tibble: 1 x 3
##    estimate statistic p.value
##       <dbl>     <dbl>   <dbl>
## 1 -9501455.     -2.65  0.0125
## # A tibble: 1 x 3
##    estimate statistic p.value
##       <dbl>     <dbl>   <dbl>
## 1 -1143084.    -0.129   0.899
## # A tibble: 1 x 3
##     estimate statistic   p.value
##        <dbl>     <dbl>     <dbl>
## 1 -10342798.     -3.98 0.0000724
## # A tibble: 1 x 3
##    estimate statistic p.value
##       <dbl>     <dbl>   <dbl>
## 1 49009455.      3.07 0.00310
## [1] "Region not printed because of few samples: 7"
## [1] "Region not printed because of few samples: 8"
## # A tibble: 1 x 3
##     estimate statistic p.value
##        <dbl>     <dbl>   <dbl>
```

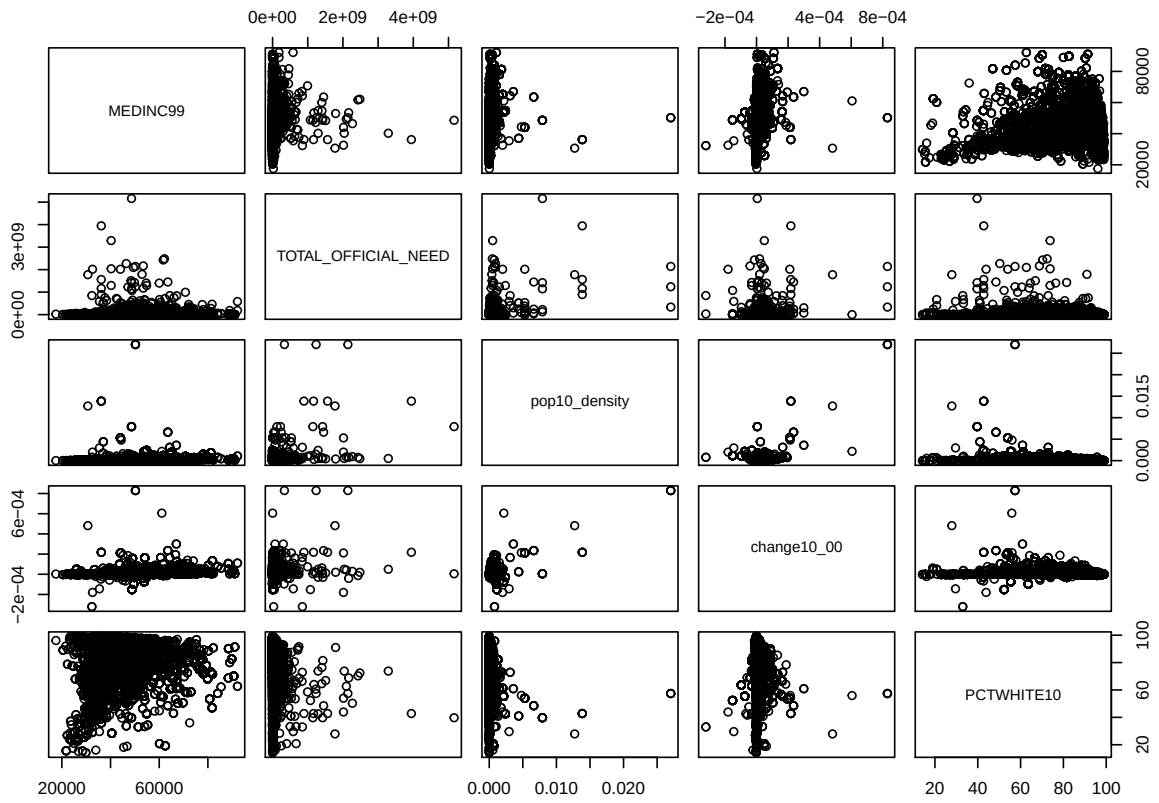
```
## 1 115576322.     0.873   0.542
## # A tibble: 1 x 3
##    estimate statistic p.value
##       <dbl>     <dbl>   <dbl>
## 1 -8802160.     -1.26   0.207
```

Some of the tests of difference of means proved inconclusive because there were too few data points to make a sound statistical comparison. However, for regions that had enough samples i.e. greater than facilities in each category, we had p - values that were above the 0.05 threshold, therefore we could not reject the null hypothesis, therefore the difference in means for those categories was intangible. We can infer that facilities in some regions that fed into a water body did not really exhibit statistically significant differences, therefore we can ignore the TMDL phenomena.

It must be noted that overall, there was a statistically significant difference in means as shown in the first t-test done on the overall data set.
In the coming part, we will see if there are some categorical variables that are associated with each other with regard to total official need. From the data given and from some analysis done earlier, I chose to explore the following variables:
- REGION
- OWNER_TYPE
- CSS
- MS4
- TMDL

The chisquare test all have p-values less that 0.05 indicating that they associations are statistically significant (see appendix for values). We can infer that some of the categorical variables have some influence on the observations, and we can make recommnedations based on some of the categorical variables

I will attempt to see if there is a correlation between need and incomes. Chosen variables are:
- MEANINC09 - MEDINC09 - POP10 - PCTWHITE10 - TOTAL_OFFICIAL_NEED



From the pairplot above, there seems to be some weak evidence of correlation among the variables chosen. However, there is a combination that seems to be positiviely correlated i.e. PCTWHITE and MEDINC09. This is to be expected since several studies on the socioeconomic status of the USA has shown that regions with high median incomes tend to have higher percentages of white residents. We can also infer from theory that these regions also have low needs, based on the scatter plot above, but we need to do paired t-tests to ascertain that these values are truly not correlated.

Some of the result from the pairwise plots are included in the diagram above. After that, we see that some of the data is weakly correlated

Section 3: Conclusion and Recommendations

From the data manipulation above, we can see that Philadelphia should implement a municipal separate sewer system. The CSS systems seem to be the ones that have the most problems, and this is to be expected since they are systems that server multiple needs i.e. stormwater, industrial sewage and home sewage, hence more demands are exerted on it. CSS systems are especially susceptible to enviromental factors e.g. a heavy downpour could damage some of the sewer systems thus placing some neighborhoods at risk.

I would also advise the mayor to avoid building a system that would empty into a waterbody. This would force the city to have to comply to extra TMDL standards set by the state. The combination of a CSS and TMDL system would force Philadelphia to invest more capital in facilities and resources that meet Pennsylvania's TMDL standards, and can withstand environmental stressors.

Appendix

Raw output for chi-square test are shown below:

```
## $<NA>
## NULL

## $chisq
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 609.5, df = 9, p-value < 2.2e-16

## $chisq
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 609.5, df = 9, p-value < 2.2e-16

## $chisq
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 626.92, df = 9, p-value < 2.2e-16

## $<NA>
## NULL

## $<NA>
## NULL

## $<NA>
## NULL

## $chisq
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 8420, df = 1, p-value < 2.2e-16

## $chisq
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 7.6843, df = 1, p-value = 0.00557
```

Raw output for some of the t-tests are indicated below:

```
##
##  Pearson's product-moment correlation
##
## data:  MEDINC99 and TOTAL_OFFICIAL_NEED
## t = 4.6783, df = 8418, p-value = 2.937e-06
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.02959600 0.07220548
## sample estimates:
##        cor
## 0.05092391
```

```
##
##  Pearson's product-moment correlation
##
## data:  TOTAL_OFFICIAL_NEED and POP10
## t = 16.355, df = 8418, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1547099 0.1961150
## sample estimates:
##       cor
## 0.1754901
```

```
##
##  Pearson's product-moment correlation
##
## data:  TOTAL_OFFICIAL_NEED and PCTWHITE10
## t = -13.935, df = 8418, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1709692 -0.1292118
## sample estimates:
##        cor
## -0.1501574
```