Data Assignment PPHA 45700

Andres Chaparro ID 12215739

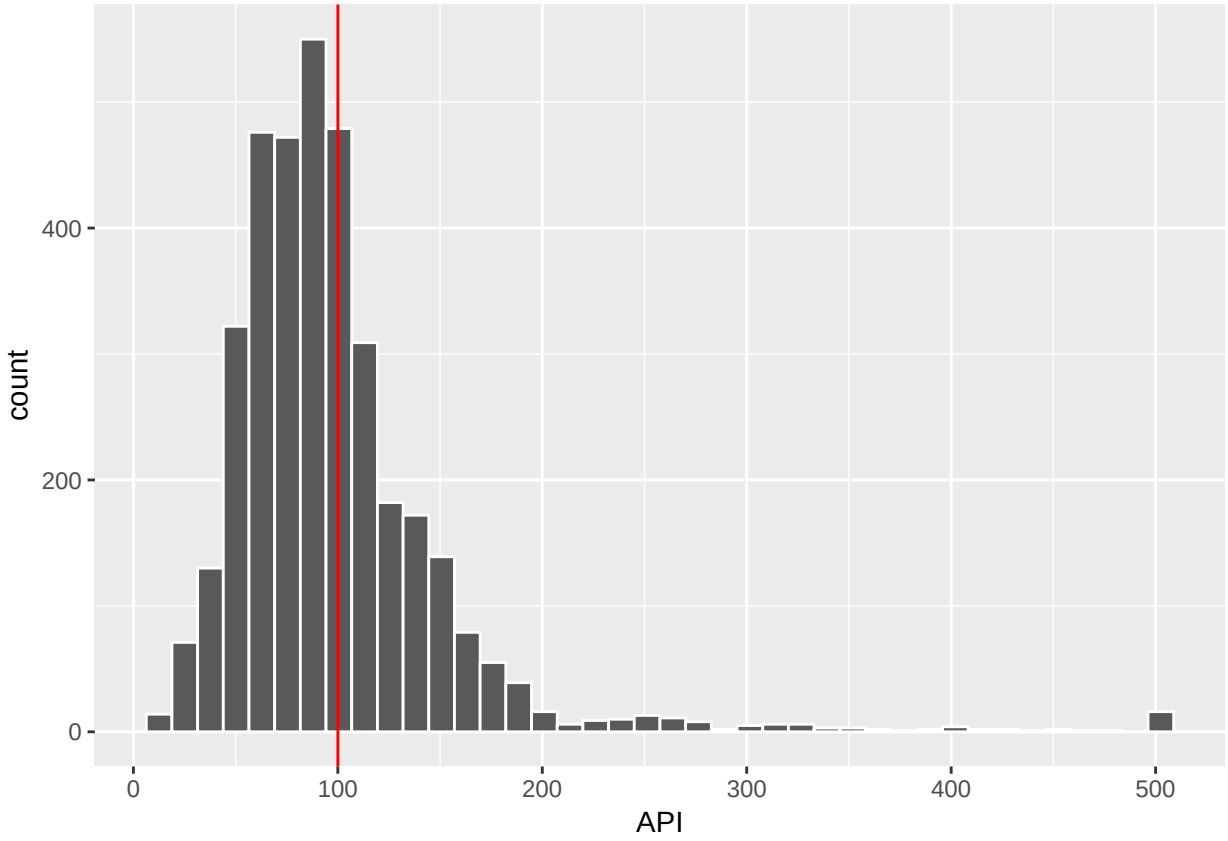
11/06/2019

1. Perfection on demand

The dataset China.Rdata provides the AQI for several years for the city of Beijing which is of naturally of particular interest. Plot the histogram of AQI for Beijing. Do you observe anything unusual? Might this affect the PhD student's planned research design, and if so why? (5 points)

The PhD student decided to use a (Sharp) Regression Discontinuity Design to explore the relationship between early promotions and employee recognition on job performance. For the methodology to be successful, the data on APIs needs to comply with the identifying assumption of the RDD. The assumption tells us that there can't be a discrete jump at the cutoff point. This means that the # of APIs are smooth around the cutoff.

We want to make sure that units don't sort around the cutoff. This is sometimes call manipulation test. We can do this by looking at the histogram and focusing on the distribution around the 100 value. Another test called covariate smoothness test looks at other predetermined variables to check they are not discontinuous at the cutoff value (to avoid picking the effect of that variable in the analysis). This second test falls outside the scope of the task at hand.



By looking at the cutoff there seems to be an unexpected agglomeration of results just below the 100 cutoff. This generates some concerns as to the validity of the Regression Discontinuity Design as described by the

PhD student, because as we've mentioned smoothness around the cutoff is the main way we try to prove the identifying assumption of RDD.

A discrete jump implies there is something else that is separating people from just below the cutoff to just above the cutoff. This is why we find the result problematic. It is the PhD student's job to determine if this jump is enough to make the RDD invalid or if the magnitude of the jump is not that great as to make it affect the validity of the design.

2. Vehicle Rationing

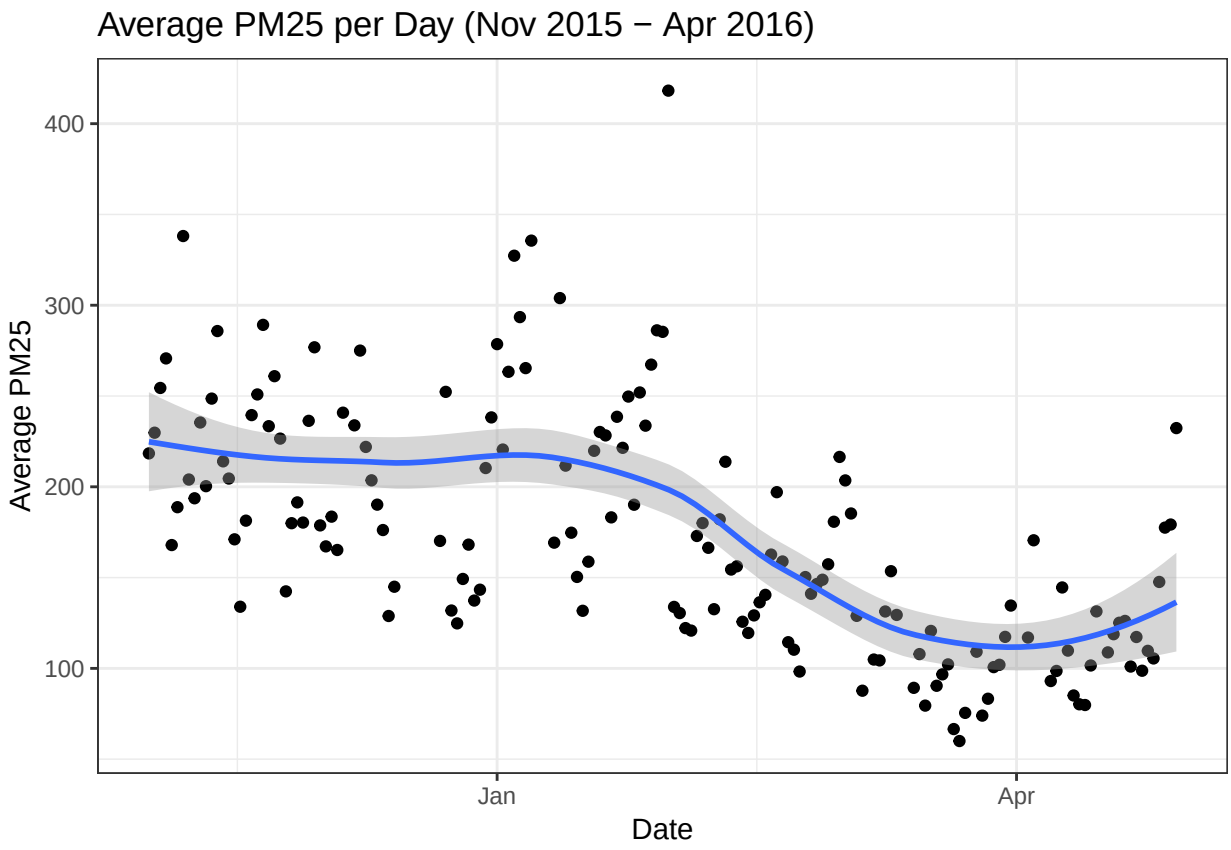
Policy of rationing driving in Dehli. Cars were classified into odd and even categories on the basis of the last digit of their license plates. It was mandated that only vehicles with odd numbered license plates could be driven on odd numbered dates and only those with even plates, on even dates. The restrictions were in force between the hours of 8 am and 8 pm for the first 15 days of January 2016. Recently they announced the policy would be repeated, starting November 4th.

Part 1

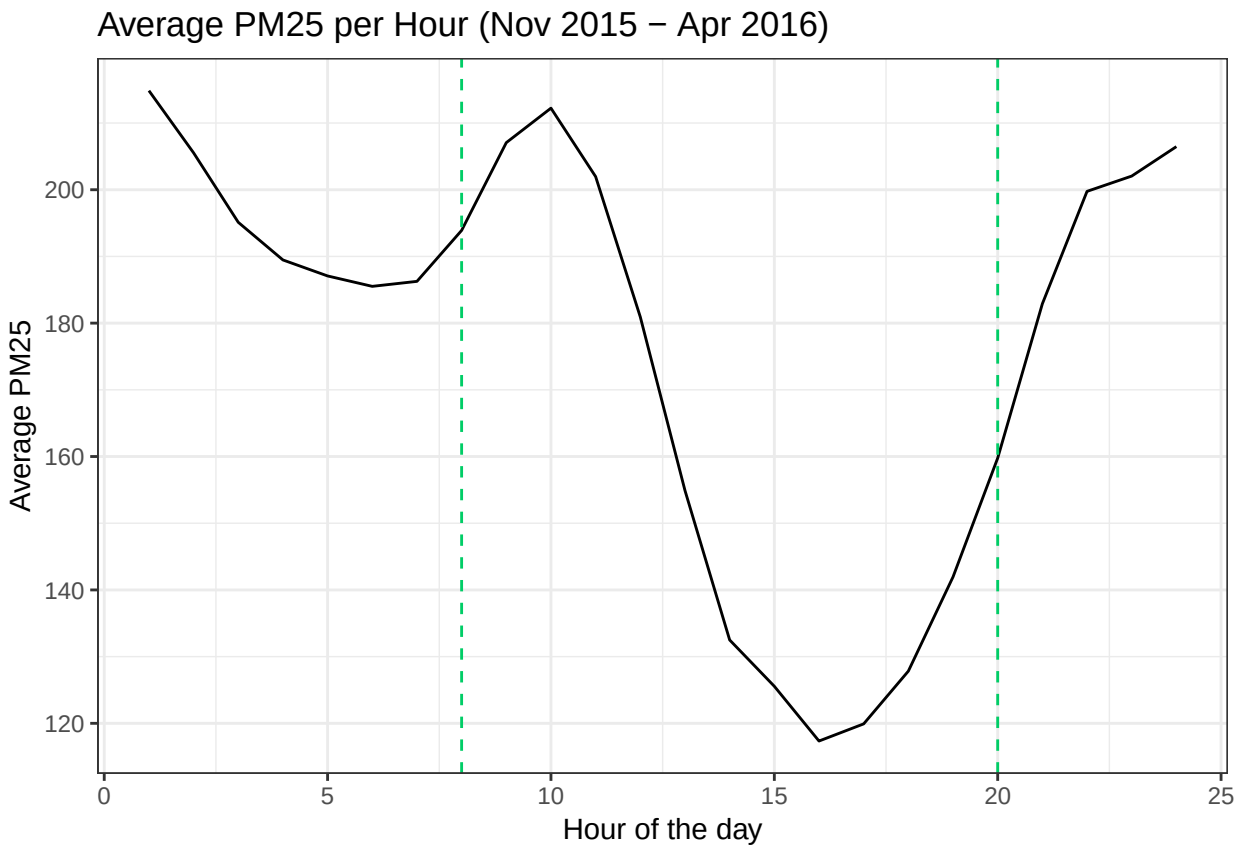
2.1.1

```
##       PM10              PM25              NO2
##  Min.   :  75.45   Min.   :  10.04   Min.   :  0.10
##  1st Qu.: 315.67   1st Qu.:  96.10   1st Qu.: 33.52
##  Median : 427.50   Median : 149.82   Median : 60.31
##  Mean   : 474.43   Mean   : 174.84   Mean   : 68.76
##  3rd Qu.: 570.73   3rd Qu.: 227.67   3rd Qu.: 95.28
##  Max.   :4265.17   Max.   : 999.99   Max.   :574.44
##  NA's   :16631                       NA's   :1904
```

2.1.2

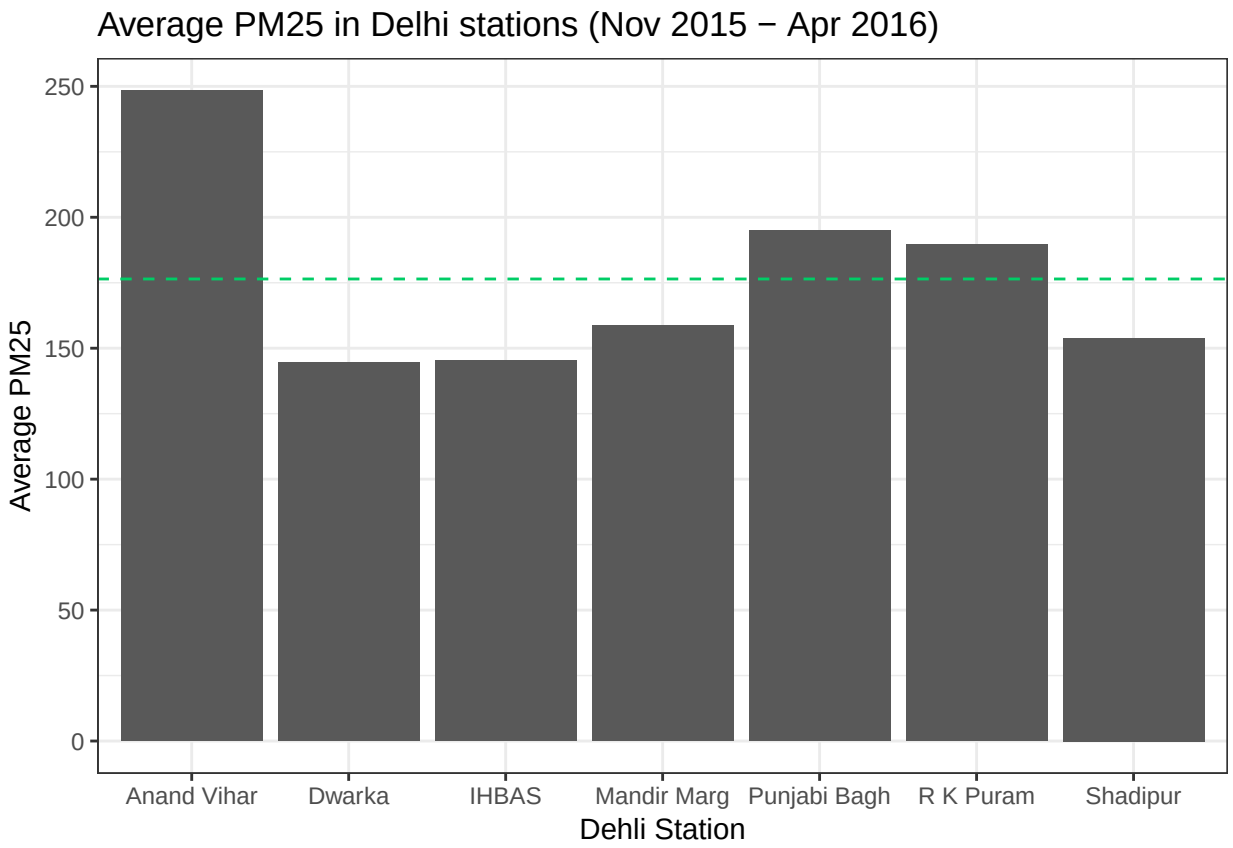

Average PM25 per Day (Nov 2015 − Apr 2016)
Average PM25
400
300
200
100
Jan
Apr
Date

2.1.3



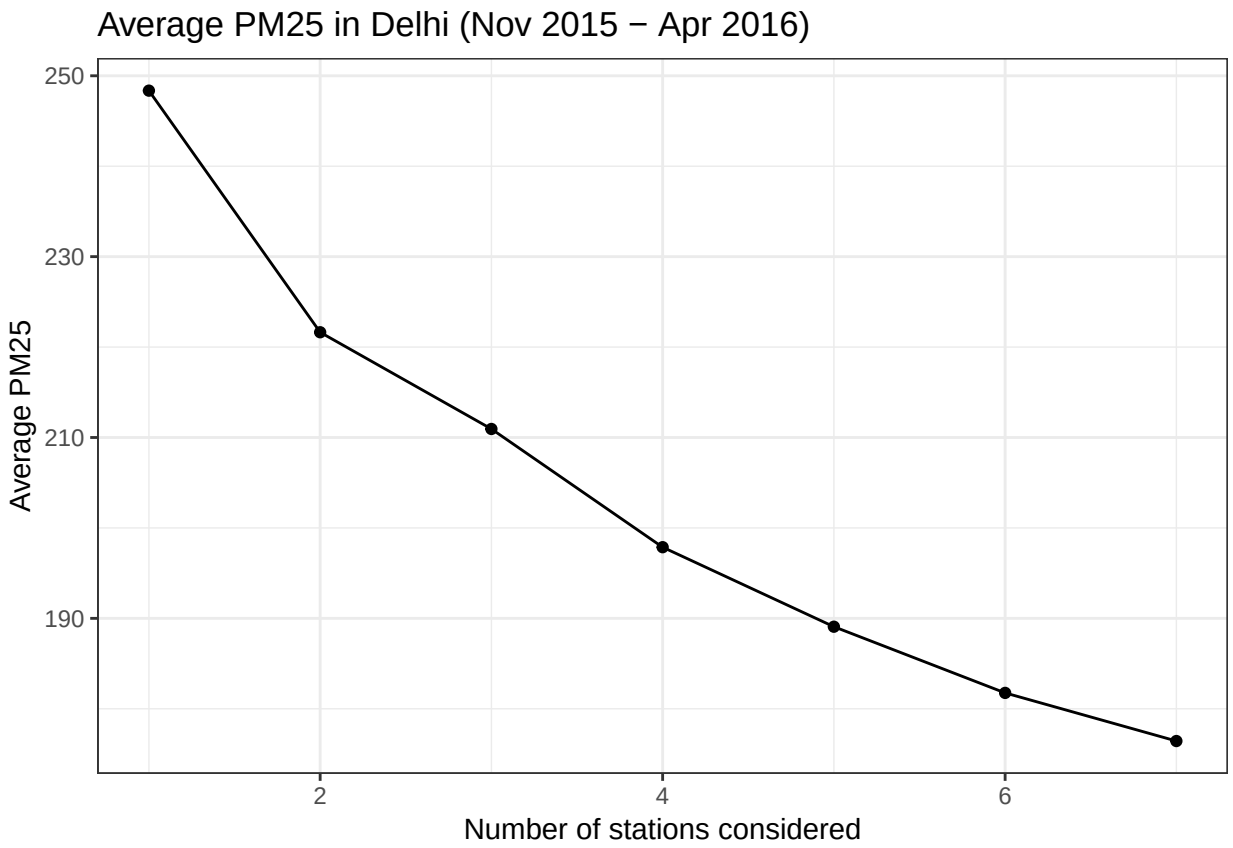
The policy is in place from 8:00 until 20:00. As we can see in the graph from 8:00 to 12:00 the pollution is at some of its highest levels of the day. However after 12:00 the pollution decreases for most of the day considerably and starts to pick up after 20:00. The highest levels of pollution actually occur before and after the policy is in place. It looks like the policy may not be addressing the main problem, although there can be other factors that may be causing PM25 pollution at night/morning and have nothing to do with driving.

2.1.4



Variations in pollution can occur due to the geographic distribution of stations. Industrial zones or zones with a lot of traffic are expected to have higher pollution levels than less populated areas, or areas without high-pollutant industries. Wind and other geological factors also influence levels of pollution.

2.1.5

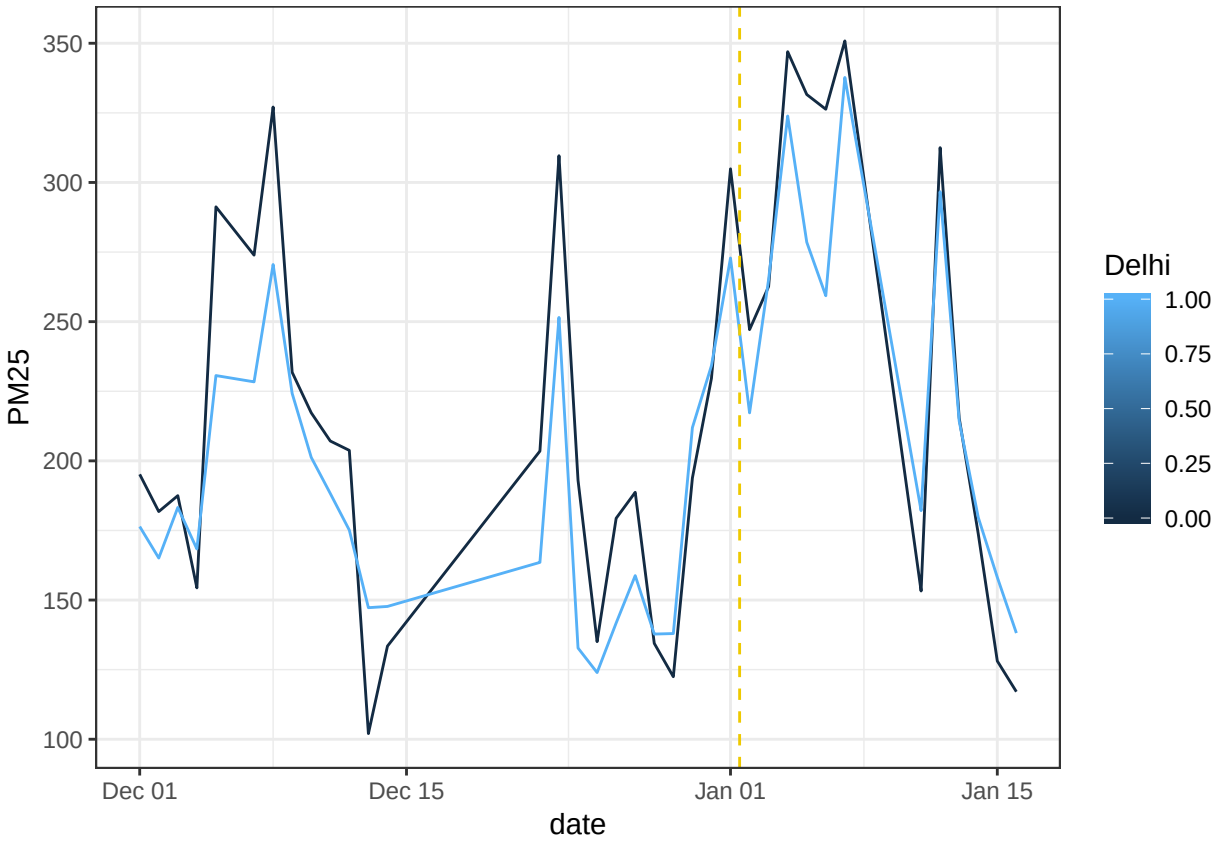


The pollution only considering the first station is 248.5 PM25 average a day for the available data. Considering all stations, the average PM25 per day goes down to 176.43. This is a 72 points decrease, or 30% decline over the initial value.

Part 2

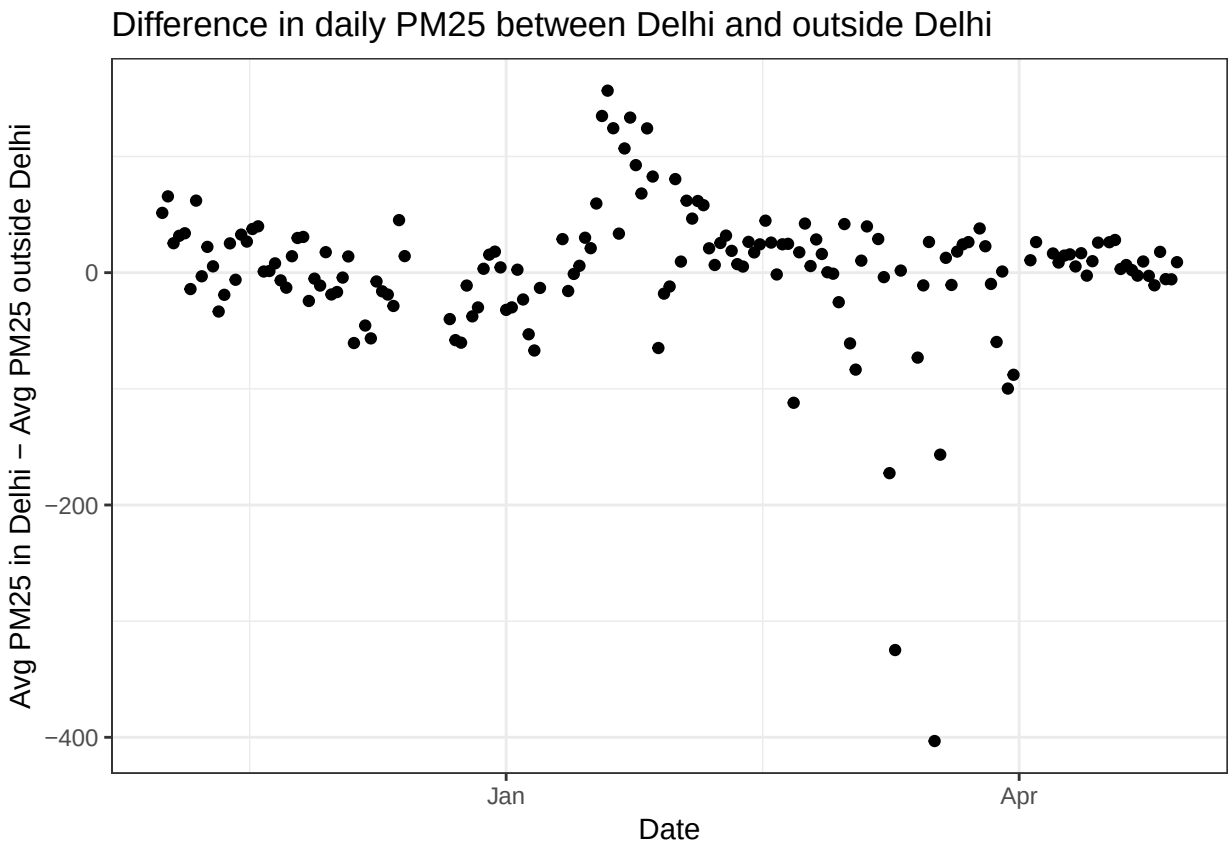
2.2.1

On a single plot show two lines - one for average pollution over time inside Delhi and the other for average pollution over time outside Delhi. Do you see any difference in pollution before and after the policy starts? Inside Delhi? Outside Delhi?



Initially the levels of PM25 went down in the first couple of dats both in Delhi and elsewhere. However the next few days we can see the highest numbers for the period. Then we have an extreme downward trend and a rebound. It is difficult to get a definitive conclusion about the impact of the policy since there is so much variation between days.

2.2.2



2.2.3

$PM25_{st} = \alpha + \gamma_s + \beta(Policy_{st}) + \epsilon_{it}$

```
##
## Call:
## lm(formula = PM25 ~ station_id + Policy, data = part2_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.60  -65.31  -15.64   55.05  339.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 194.1974     6.0432  32.135  < 2e-16 ***
## station_id   -3.2366     0.7334  -4.413 1.13e-05 ***
## Policy       71.3386     9.8217   7.263 7.58e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.17 on 1002 degrees of freedom
## Multiple R-squared:  0.06774,    Adjusted R-squared:  0.06588
## F-statistic:  36.4 on 2 and 1002 DF,  p-value: 5.474e-16
```

We find a positive relationship between Policy in place and PM25 levels. This is to say that on the time that

the policy was in place we saw an increase in PM25. The estimates are significant under 99% confidence interval.

The coefficienct for *Policy* can be explained by other factors. One can be that it was because of the expected increase in pollution that the policy was put in place. There can be other variables that we are ommitting, however it doesn't paint a positive picture for the policy.

2.2.4

$PM25_{st} = \alpha + \gamma_s + \beta(Policy_{st} * time) + \epsilon_{it}$

```
##
## Call:
## lm(formula = PM25 ~ station_id + Policy * time, data = part2_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -200.83  -55.30   -6.83   42.25  361.21
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 257.02660    6.67953  38.480  < 2e-16 ***
## station_id   -3.03704    0.65132  -4.663 3.54e-06 ***
## Policy      531.13292  113.71946   4.671 3.41e-06 ***
## time         -0.72021    0.04559 -15.799  < 2e-16 ***
## Policy:time  -6.87081    1.64198  -4.184 3.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.51 on 1000 degrees of freedom
## Multiple R-squared:  0.2665, Adjusted R-squared:  0.2636
## F-statistic: 90.83 on 4 and 1000 DF,  p-value: < 2.2e-16
```

We still have all estimates significant at the 99%. In this case the *Policy* coefficient is still positive. Nevertheless, both time and $Policy * time$ are negative. The coefficient of *Policy* is the effect of the policy only when time is zero. It is kind of difficult to picture. The coefficient of *time* is the effect of time on PM25 when Policy is zero. This would measure how much is the pollution increasing or decreasing through time. Finally the coefficient of $Policy * time$ is the effect of time and Policy together. So the effect of time when the policy is active. We could argue that as time passes when the policy is active, more people adapt and reduce car usage, thus reducing PM25. This is a possible explanation but not by any means the only thing we could interpret from this data.

2.2.5

$PM25_{st} = \alpha + \gamma_s + \beta(Delhi * Policy_{st}) + \epsilon_{it}$

```
##
## Call:
## lm(formula = PM25 ~ time + Policy * Delhi, data = part2_5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.11  -54.98   -9.96   39.26  747.70
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  232.17485    7.29465  31.828   <2e-16 ***
## time          -0.70017    0.04467 -15.676   <2e-16 ***
## Policy        47.86168   21.38756   2.238   0.0254 *
## Delhi          0.97332    6.30047   0.154   0.8773
## Policy:Delhi   9.70604   23.31537   0.416   0.6773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.3 on 1228 degrees of freedom
## Multiple R-squared:  0.2062, Adjusted R-squared:  0.2036
## F-statistic: 79.75 on 4 and 1228 DF,  p-value: < 2.2e-16
```

In this case the coefficients from $Policy * Delhi$ and $Delhi$ are not statistically significant at a 95% Confidence Interval. This is to say that we can't really draw conclusions from this regression. Because the policy is in Delhi, the value we would be interested is $Policy * Delhi$ because that is the effect of the policy on Delhi. The policy shouldn't have an big effect outside of Delhi, and we are not looking at differences between Delhi and the outside so our attention should be on $Policy * Delhi$.

2.2.6

$PM25_{sht} = \alpha + \gamma_t + \omega s + \theta h + \beta(Delhi * Policy_{sht}) + \epsilon_{sht}$

```
##
## Call:
## lm(formula = PM25 ~ date + station_id + Hour + Policy * Delhi,
##     data = part2_6_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239.33  -69.91  -16.09   50.10  887.57
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.185e+04  2.123e+02  55.842   <2e-16 ***
## date         -8.015e-06  1.460e-07 -54.911   <2e-16 ***
## station_id   -3.754e+00  1.938e-01 -19.375   <2e-16 ***
## Hour         -1.749e+00  9.752e-02 -17.936   <2e-16 ***
## Policy        5.686e+01  6.468e+00   8.790   <2e-16 ***
## Delhi         1.937e+01  1.875e+00  10.327   <2e-16 ***
## Policy:Delhi -8.061e+00  7.061e+00  -1.142    0.254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.1 on 22982 degrees of freedom
## Multiple R-squared:  0.164,  Adjusted R-squared:  0.1637
## F-statistic: 751.2 on 6 and 22982 DF,  p-value: < 2.2e-16
```

Every coefficient is statistically significant except the $Polcy * Delhi$ interaction term. The result in this case makes a case for a really small decrease in pollution due to the policy for stations in Delhi. However, the result is too small and not statistically significant so we can't draw many conclussions.