Lastname_ENV872_Project.pdf

Ying Liu, Yingchi Cheung, Xuening Tang

2022-12-14

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
getwd()
```

```
## [1] "C:/Users/Alina/Desktop/DUKE_22FALL/872/YingLiu-YingchiCheung-XueningTang"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(cowplot)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following object is masked from 'package:cowplot':
##
##     stamp
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
mytheme <- theme_classic(base_size = 10) + theme(axis.text = element_text(color = "black"),
    legend.position = "bottom")
theme_set(mytheme)
```

1. Introduction

Water shortages around US due to climate change is becoming worse and reported by several studies. Also we are in the situation that two of the largest reservoirs in America, which provide water and electricity tomillions, are in danger of reaching 'dead pool status. With the rapid urbanization in cities around the world and in US, the amount of population is rapidly increasing and the imperviousness of urban surface is increasing as well, which could raise the need for water withdrawal from the lake bodies or reservoirs and the amount of wastewater discharge from daily life.

At the same time, the intemperate discharge of wastewater in US is becoming a critical issue so in this way, we would like to investigate two main questions in our report: 1)Is there any clear correlation between the monthly average water withdrawal and the monthly average wastewater discharge? 2)Is population a key influencing factor to water withdrawal and wastewater discharge over the years?

2. Data

2.1 Read Datasets

To solve these problems, we gathered three main datasets from the website of NC DEQ Division of Water Resources-Local Water Supply Planning (https://www.ncwater.org/WUDC/app/LWSP/search.php) for monthly water withdrawal data and monthly wastewater discharge data; from the website of Data Commons (https://datacommons.org/place/geoId/3719000?category=Demographics#Population) for population data in Durham county from 2012-2021.

```
population <- read.csv("./data/Population_new.csv", stringsAsFactors = TRUE)

water.withdraw <- read.csv("./data/EDA final data_water.csv",
    stringsAsFactors = TRUE)

water.discharge <- read.csv("./data/EDA_final project_wastewater_discharge.csv",
    stringsAsFactors = TRUE)
```

2.2 Data Wrangling

In the data wrangling part, we first filtered population data we need for ten years and adjusted the formats of every column of data into the one more convenient for later steps. Then we joined the table for withdrawal data, discharge data and population data by year and added the column of month to distinguish. Also, we created one dataset which summmarized the mean discharge amount and the mean withdrawal amount in each month among years.

```
# 1.First step eliminate the extra column in data set and
# convert it into format we need.
durh_pop <- population %>%
    filter(Year == 2012 | Year == 2013 | Year == 2014 | Year ==
        2015 | Year == 2016 | Year == 2017 | Year == 2018 | Year ==
        2019 | Year == 2020 | Year == 2021)

# 2.Next join the tables together as our main data frame to
```

```
# conduct analysis.
water_final <- left_join(water.discharge, water.withdraw, by = c("Year",
    "Month"))

pop_water_final <- left_join(durh_pop, water_final)
```

```
## Joining, by = "Year"
```

```
pop_water_final$Population <- as.numeric(pop_water_final$Population)

pop_water_summaries <- pop_water_final %>%
    group_by(Year, Population) %>%
    summarise(mean.discharge = mean(Monthly.Discharge.Wastewater),
        sd.discharge = sd(Monthly.Discharge.Wastewater), mean.withdraw = mean(Monthly.Water.Withdraw),
        sd.withdraw = sd(Monthly.Water.Withdraw))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

In the data wrangling part, first we read the population data; the monthly water withdrawal data and the monthly wastewater discharge data into our file.

```
write.csv(pop_water_final, row.names = FALSE, file = "./data/Durham_water&population_2012-2021.csv")
write.csv(pop_water_summaries, row.names = FALSE, file = "./data/Durham_county_water_withdraw&diacharge_
```

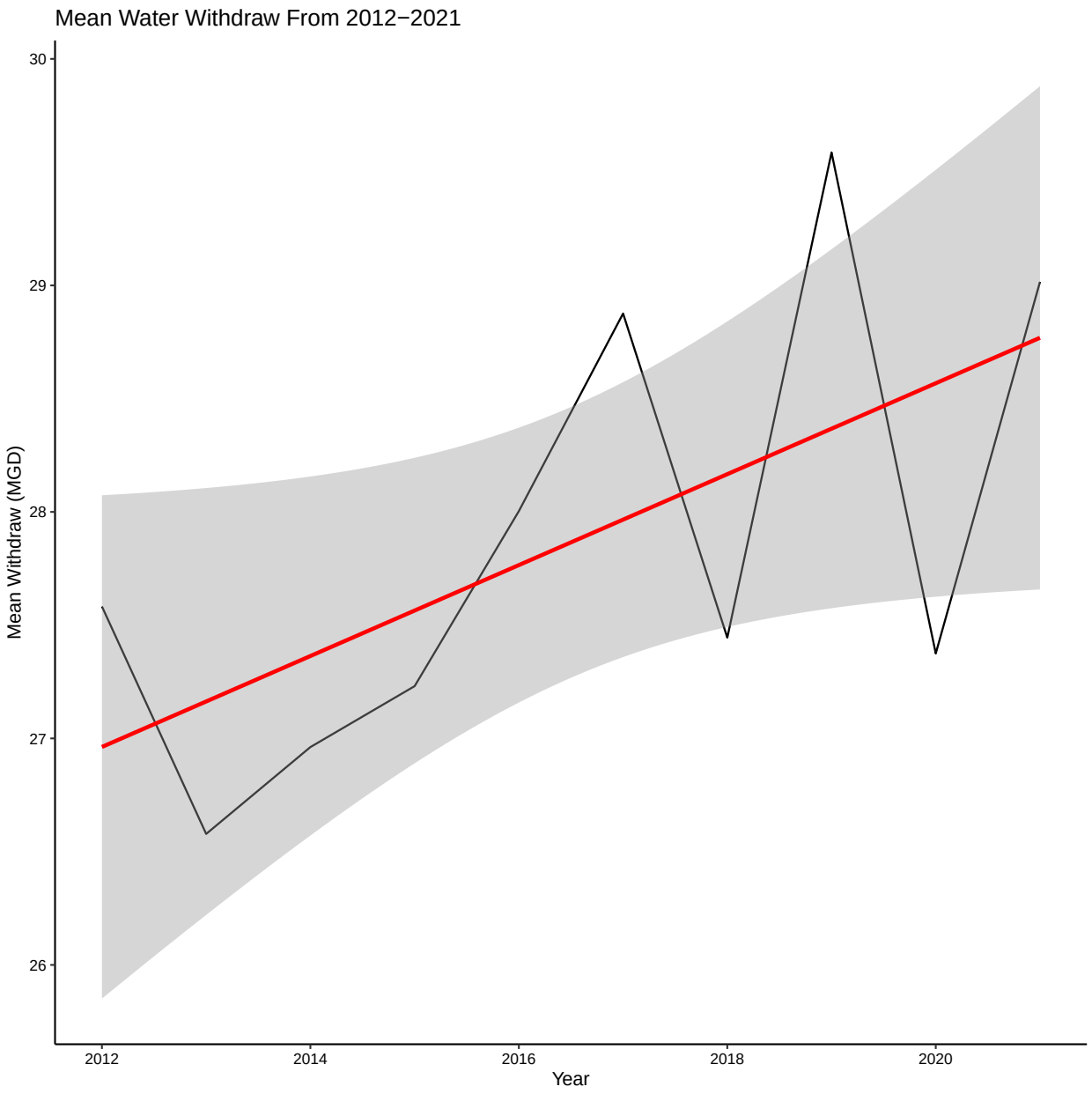
3. Data Visualization

3.1 Line Plot

We created some data visualizations to make the data that we collected easy to understand. Frist, we made line plots for the average annual wastewater discharge, water withdrawal, and population for Durham from 2012 to 2021 using ggplot. A line plot is used for continuous variables over time. Over the course of 10 years, we can see in Figure 1 below that all three variables have an overall increasing trend.

```
discharge_linechart <- ggplot(pop_water_summaries, aes(x = Year,
    y = mean.discharge)) + geom_line() + labs(title = "Mean Wastewater Discharge From 2012-2021",
    y = "Mean Discharge (MGD)", x = "Year") + labs(title = "Figure 1. Mean Wastewater Discharge",
    subtitle = "From 2012-2021", y = "Mean Discharge (MGD)",
    x = "Year") + geom_smooth(method = lm)

withdraw_linechart <- ggplot(pop_water_summaries, aes(x = Year,
    y = mean.withdraw)) + geom_line() + labs(title = "Mean Water Withdraw From 2012-2021",
    y = "Mean Withdraw (MGD)", x = "Year") + geom_smooth(method = lm,
    color = "red")
plot(withdraw_linechart)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

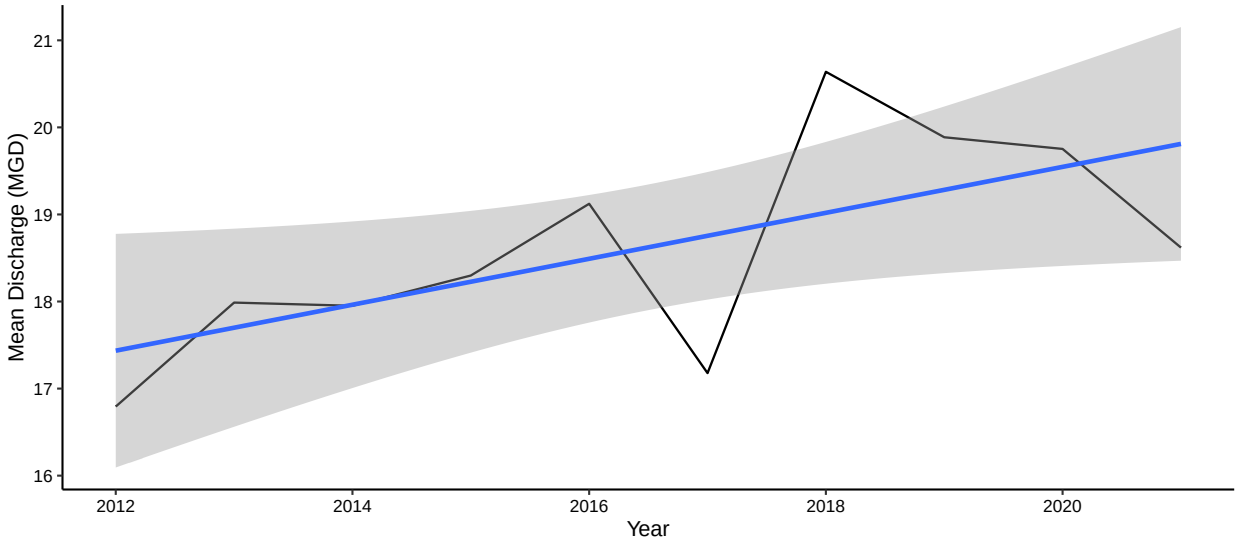



```
plot_grid(discharge_linechart, withdraw_linechart, ncol = 1,
    align = "v", axis = "l")
```

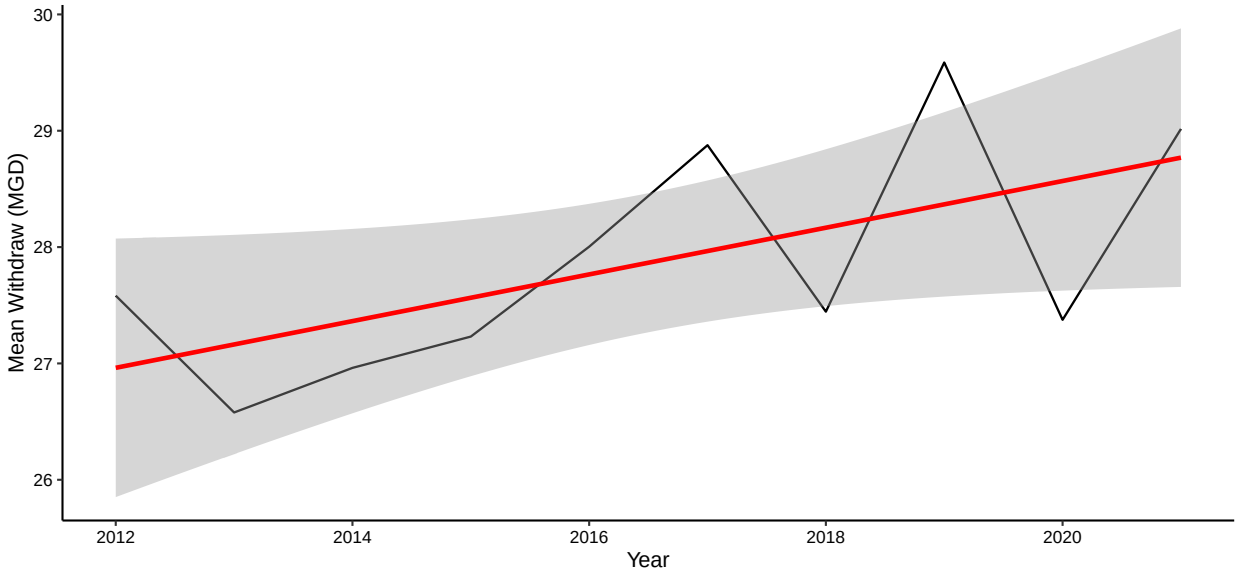
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

Figure 1. Mean Wastewater Discharge

From 2012−2021



Mean Water Withdraw From 2012−2021

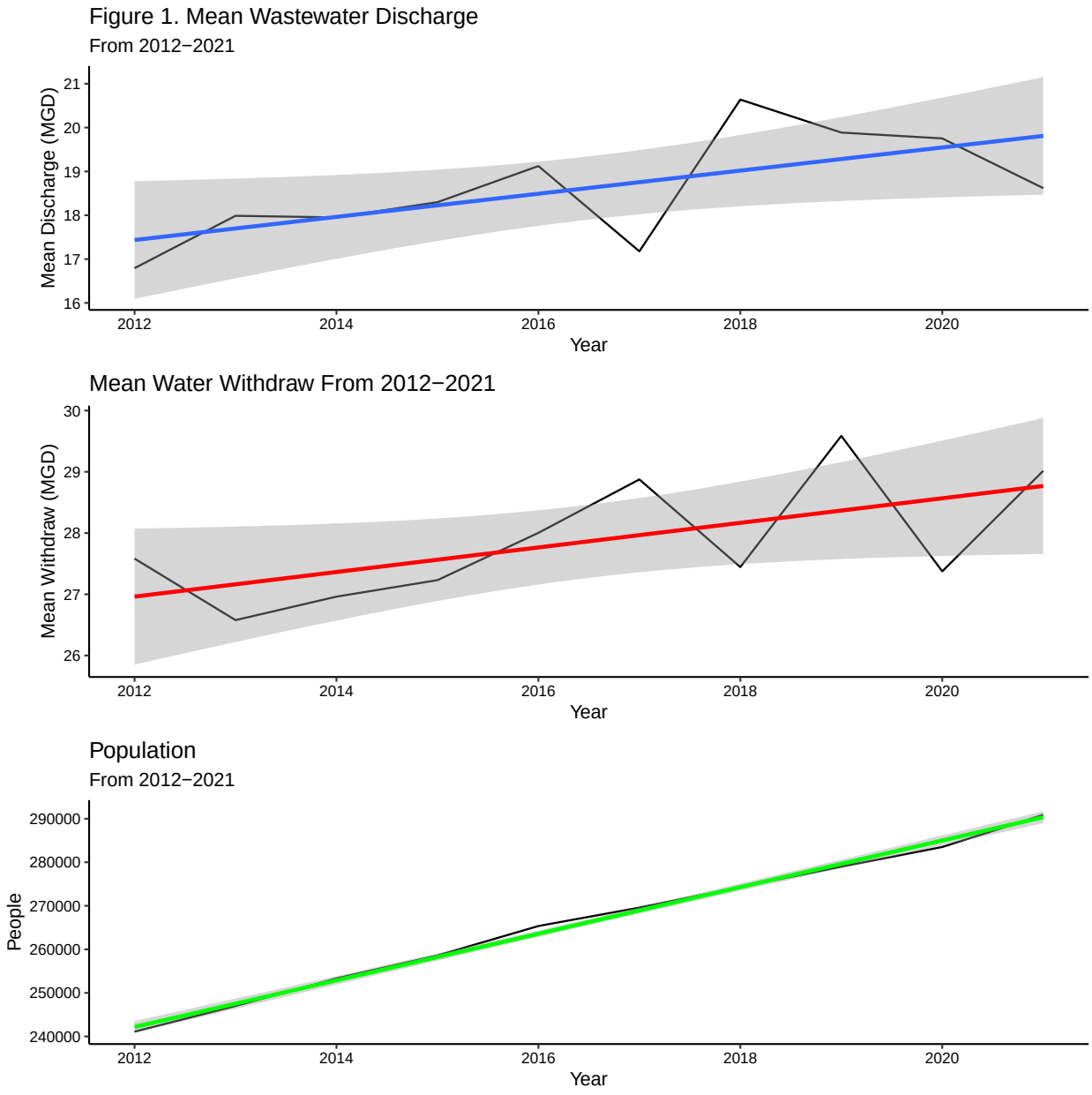



```
labs(title = "Mean Water Withdrawal", subtitle = "From 2012-2021",
    y = "Mean Withdrawal (MGD)", x = "Year") + geom_smooth(method = lm,
    color = "red")
```

```
## NULL
```

```
pop_linechart <- ggplot(pop_water_summaries, aes(x = Year, y = Population)) +
    geom_line() + labs(title = "Population", subtitle = "From 2012-2021",
    y = "People", x = "Year") + geom_smooth(method = lm, color = "green")

plot_grid(discharge_linechart, withdraw_linechart, pop_linechart,
    ncol = 1, align = "v", axis = "l")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



3.2 Box Plot

To examine the seasonal change of wastewater discharge and water withdrawal and the relationship between them, we made box plots (Figure 2) to see if there's any significant difference between each month in the last ten years for Durham. The monthly wastewater discharges from July to November are significantly different from the wastewater discharges from January to April. Therefore, during most of the summer and fall time, wastewater discharges are significantly lesser wastewater discharge than in most of the spring and winter time. Moreover, water withdrawals from January to March and November are significantly lower than from April to October. The water withdrawal in December is significantly lower than in all the other months,

except in January. We conclude that summer's and most of the fall's water withdrawals are the highest. Also, winter has the lowest water withdrawal.

We were expecting water withdrawal and wastewater discharge to be proportional before we were making the data analysis. Using common sense that high water withdrawal means high water demand, which results in high wastewater discharge. However, the box plots show differently. The water withdrawals and wastewater discharges in Durham from 2012 to 2021 have an inverse relationship. The water withdrawals during summer and fall are the highest, but the wastewater discharges are the lowest.

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.2.2
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 4.2.2
```

```
##
## Attaching package: 'reshape'
```

```
## The following objects are masked from 'package:reshape2':
##
##     colsplit, melt, recast
```

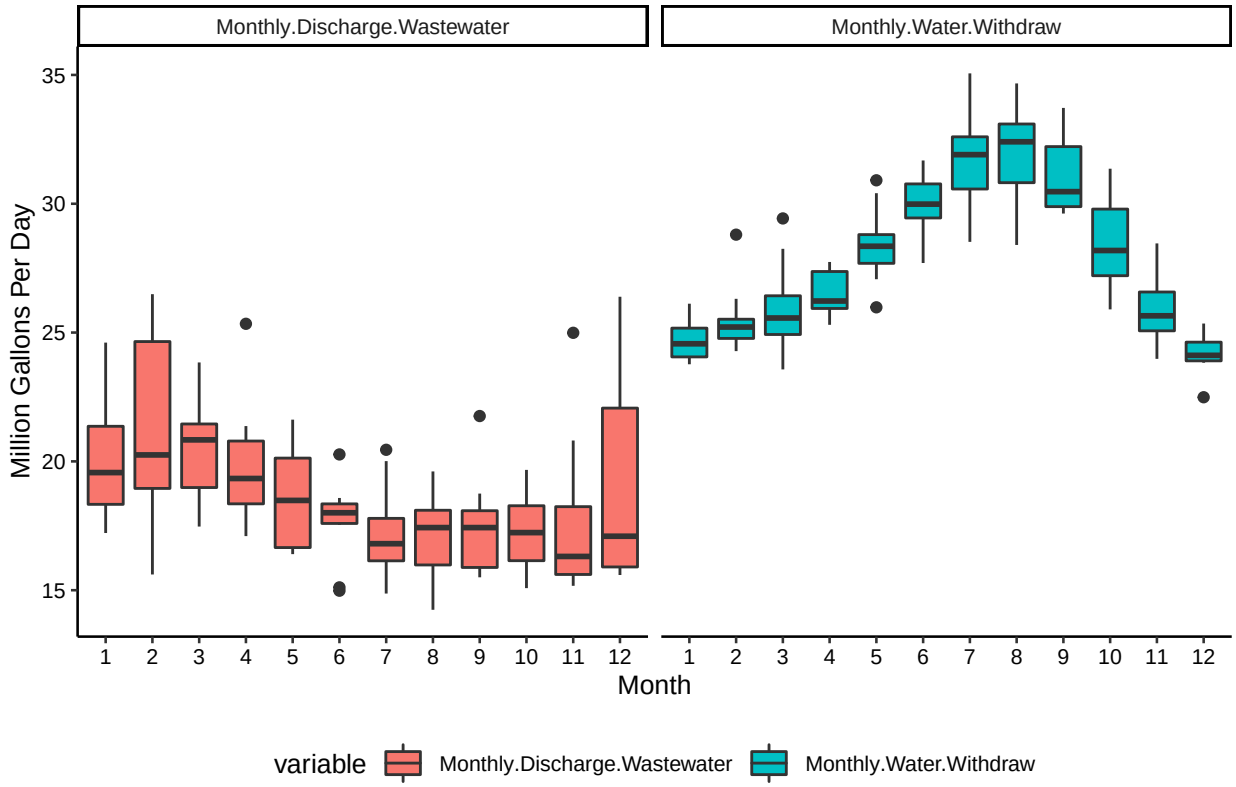
```
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
## The following object is masked from 'package:cowplot':
##
##     stamp
```

```
## The following object is masked from 'package:dplyr':
##
##     rename

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

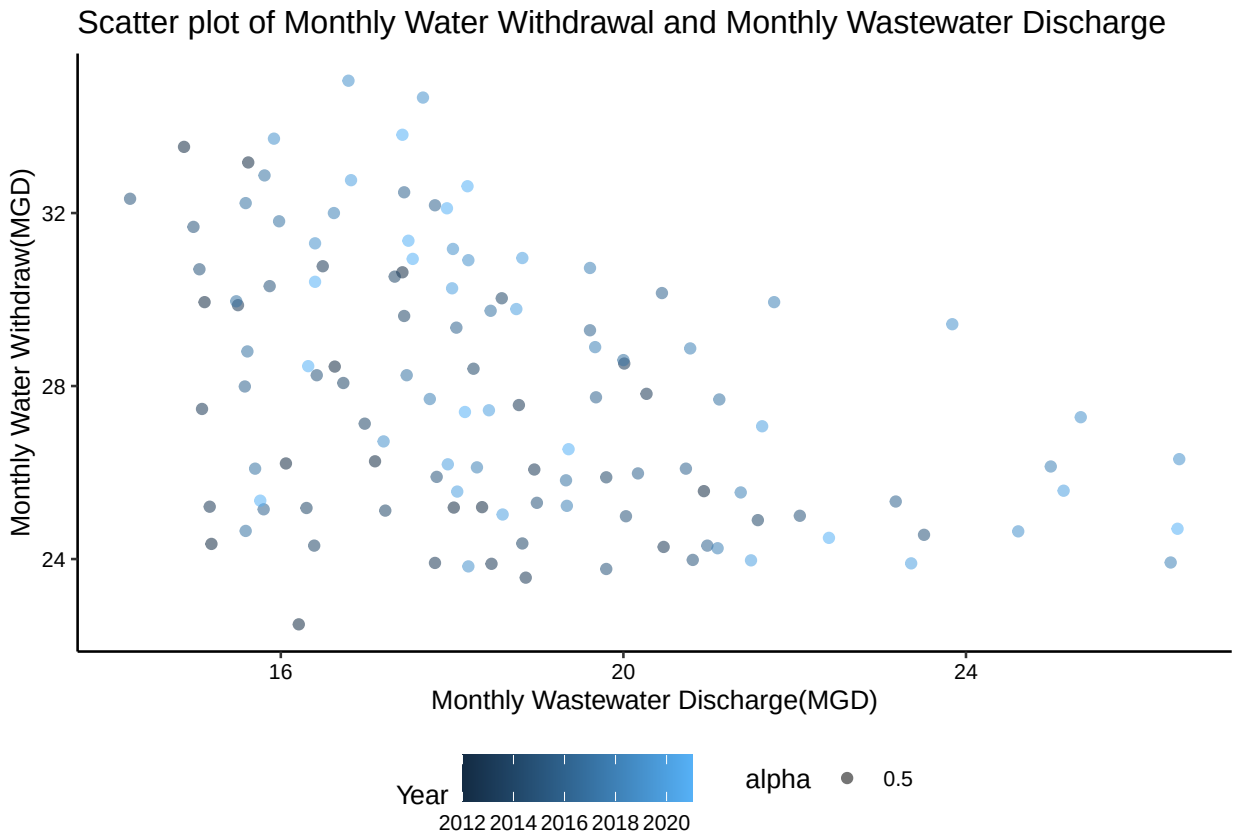
```
discharge_withdraw_grouped <- melt(water_final, id = c("Year",
    "Month"))

Wastewater_box <- ggplot(discharge_withdraw_grouped, aes(x = Month,
    y = value, group = Month, fill = variable))

Wastewater_box <- ggplot(discharge_withdraw_grouped, aes(x = as.factor(Month),
    y = value, group = as.factor(Month), fill = variable)) +
    geom_boxplot() + facet_wrap(~variable) + labs(title = "Figure 2.",
    x = "Month", y = "Million Gallons Per Day") + mytheme

print(Wastewater_box)
```

Figure 2.



>3.3 Data Analysis > >3.3.1 Methodology To explore relationship between population,monthly water withdraw,and monthly wastewater discharge,we first did a scatterplot between each two of the variablesm to observe a trend. Then created linear model, examined results and visualize results. > >3.3.2 Water Dicharge and Water withdrawal Scatter plot of monthly water discharge and withdrawal exhibited a negative trend between the two despite some cluttering. Monthly waste water discharge increased as monthly water withdrawal decreased.

```
# draw scatterplot to see if a trend exsist
discharge_withdraw <- ggplot(pop_water_final, aes(x = Monthly.Discharge.Wastewater,
    y = Monthly.Water.Withdraw)) + geom_point(aes(color = Year,
    alpha = 0.5)) + labs(title = "Scatter plot of Monthly Water Withdrawal and Monthly Wastewater Discha
    x = "Monthly Wastewater Discharge(MGD)", y = "Monthly Water Withdraw(MGD)") +
    mytheme
print(discharge_withdraw)
```



>We used linear regression to determine if there is significant relationship. P-value is close to zero which indicates that there is high signicifance between monthly waste water discharge and water withdrawal.The distribution of the residuals is not strongly symmetrical, meaning that predicted water discharge points fall far away from the actual observed points.Residual standard error is 2.5, given that mean monthly water withdrawal is 29.18,the precentage error is 8.57%. These all ponits to a high revelance between water dischagre and water withdrawal.However,the multiple R-squared shows that only 17.5% precent of the variance can be explained by water discharge.

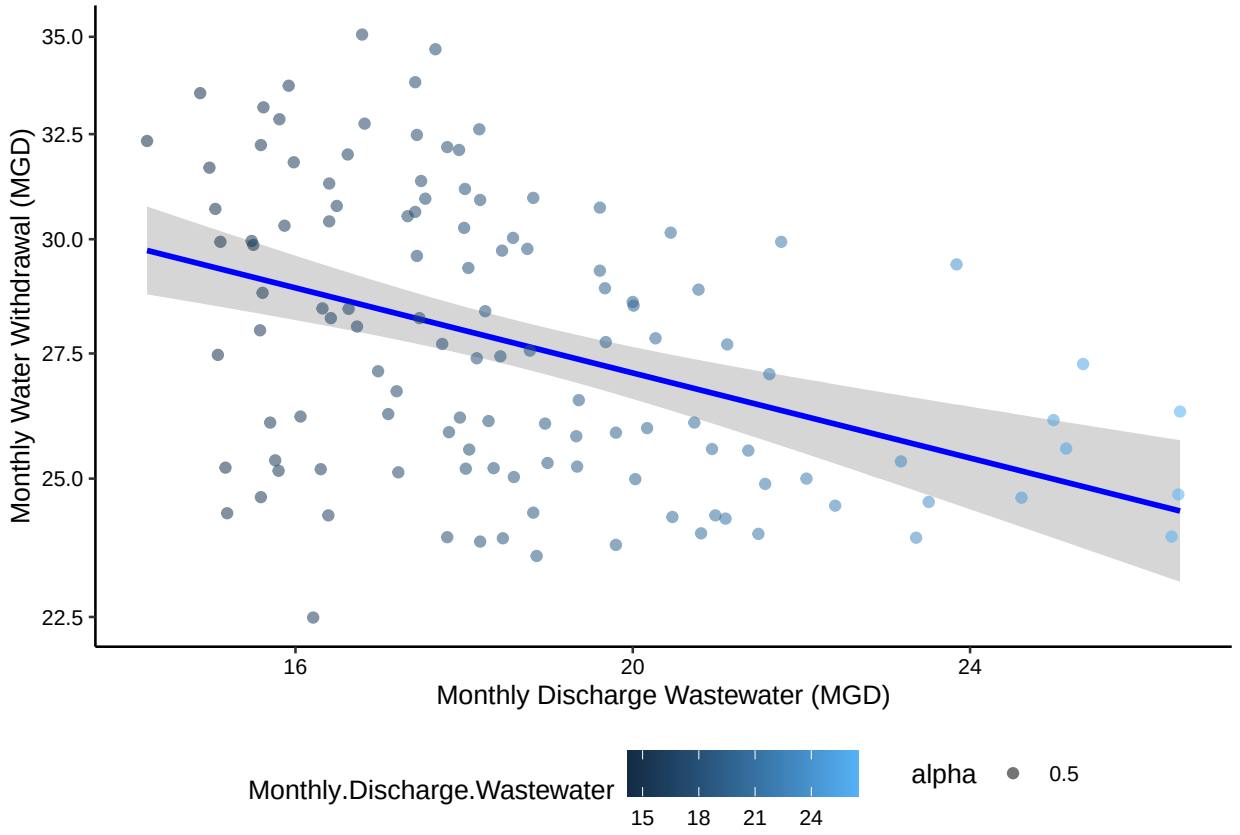
```
# Format the lm() function
discharge_withdraw <- lm(data = pop_water_final, Monthly.Discharge.Wastewater ~
    Monthly.Water.Withdraw)
summary(discharge_withdraw)
```

```
##
## Call:
## lm(formula = Monthly.Discharge.Wastewater ~ Monthly.Water.Withdraw,
##     data = pop_water_final)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7646 -1.6944 -0.2257  1.3752  7.2780
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             29.18001    2.12915  13.705  < 2e-16 ***
## Monthly.Water.Withdraw -0.37887    0.07597  -4.987 2.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 118 degrees of freedom
## Multiple R-squared:  0.1741, Adjusted R-squared:  0.1671
## F-statistic: 24.87 on 1 and 118 DF,  p-value: 2.127e-06
```

```
# Add a line and standard error for the linear regression
discharge_withdraw_regression <- ggplot(pop_water_final, aes(x = Monthly.Discharge.Wastewater,
    y = Monthly.Water.Withdraw)) + geom_smooth(method = "lm",
    color = "blue") + scale_y_log10() + geom_point(aes(color = Monthly.Discharge.Wastewater,
    alpha = 0.5)) + labs(x = "Monthly Discharge Wastewater (MGD)",
    y = "Monthly Water Withdrawal (MGD)") + mytheme

print(discharge_withdraw_regression)
```

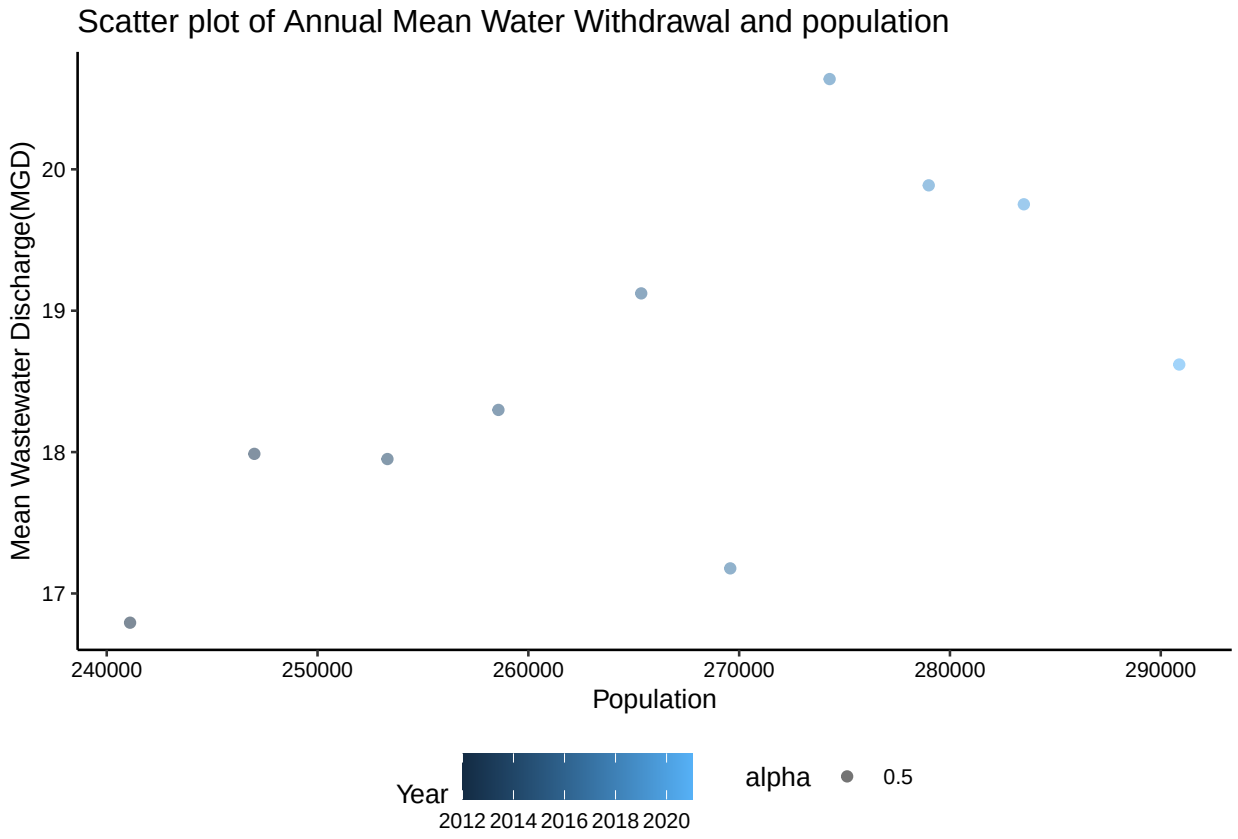
```
## `geom_smooth()` using formula 'y ~ x'
```



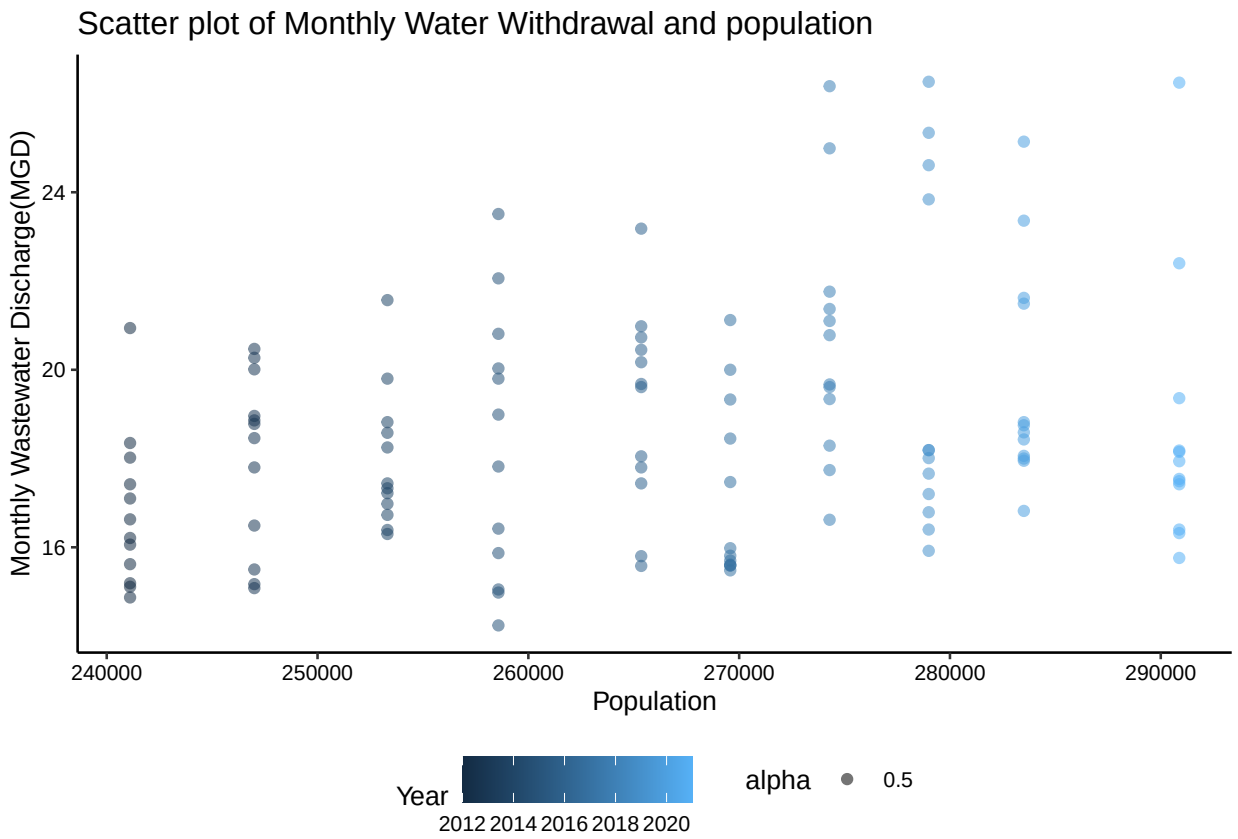
>3.3.3 Population and Water withdrawal > >Due to the time-scale difference in population and water

withdral data, this analysis uses two scatter plots:annual mean water withdrawal and population, monlth water withdrawal and population. Both shows a positive trend between the two. Monthly water withdrawal increase as population expanded during the last ten years.

```
# scatterplot of annual population and annual mean
# discharge
population_discharge <- ggplot(pop_water_summaries, aes(x = Population,
    y = mean.discharge)) + geom_point(aes(color = Year, alpha = 0.5)) +
    labs(title = "Scatter plot of Annual Mean Water Withdrawal and population",
        x = "Population", y = "Mean Wastewater Discharge(MGD)") +
    mytheme
print(population_discharge)
```




```
# scatterplot of annual population and monthly mean
# discharge
population_discharge_scatter <- ggplot(pop_water_final, aes(x = Population,
    y = Monthly.Discharge.Wastewater)) + geom_point(aes(color = Year,
    alpha = 0.5)) + labs(title = "Scatter plot of Monthly Water Withdrawal and population",
    x = "Population", y = "Monthly Wastewater Discharge(MGD)") +
    mytheme
print(population_discharge_scatter)
```



>P-value is close to zero which indicates a signicifant relationship between monthly water discharge and population.The distribution of the residuals is fairly symmetrical, meaning that predicted water discharge points fall near the actual observed points.Residual standard error is 14930, given that mean monthly water discharge is 231371,the precentage error is 6.45%. Thus, we can conclude that Durham population growth is an significant influencing factor on monthly water discharge over the last ten years .

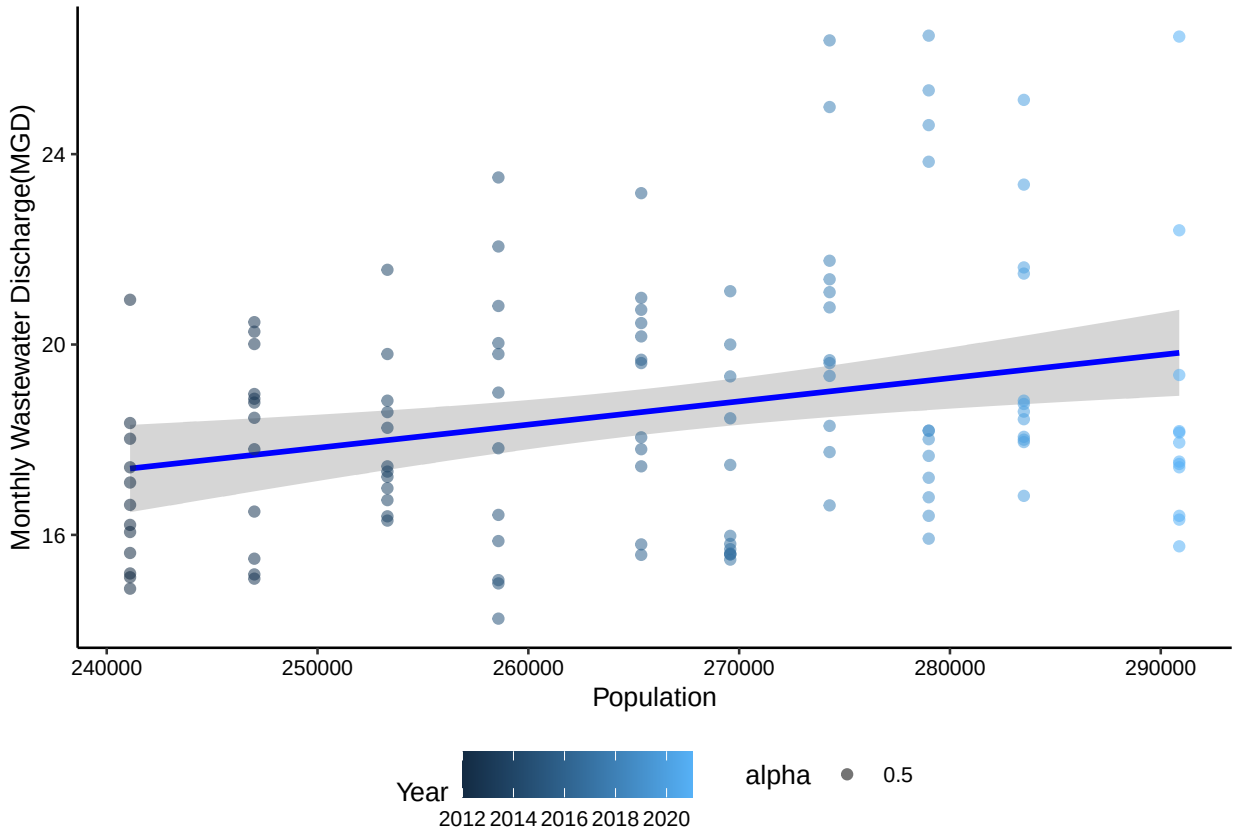
```
# Format the lm() function

population_discharge <- lm(data = pop_water_final, Population ~
    Monthly.Discharge.Wastewater)
summary(population_discharge)
```

```
##
## Call:
## lm(formula = Population ~ Monthly.Discharge.Wastewater, data = pop_water_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28748.4 -12494.5    449.3  11463.6  29057.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   237371.9     9387.1   25.29  < 2e-16 ***
## Monthly.Discharge.Wastewater   1551.3      498.7    3.11  0.00234 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 14930 on 118 degrees of freedom
## Multiple R-squared:  0.07578,    Adjusted R-squared:  0.06794
## F-statistic: 9.675 on 1 and 118 DF,  p-value: 0.002343
```

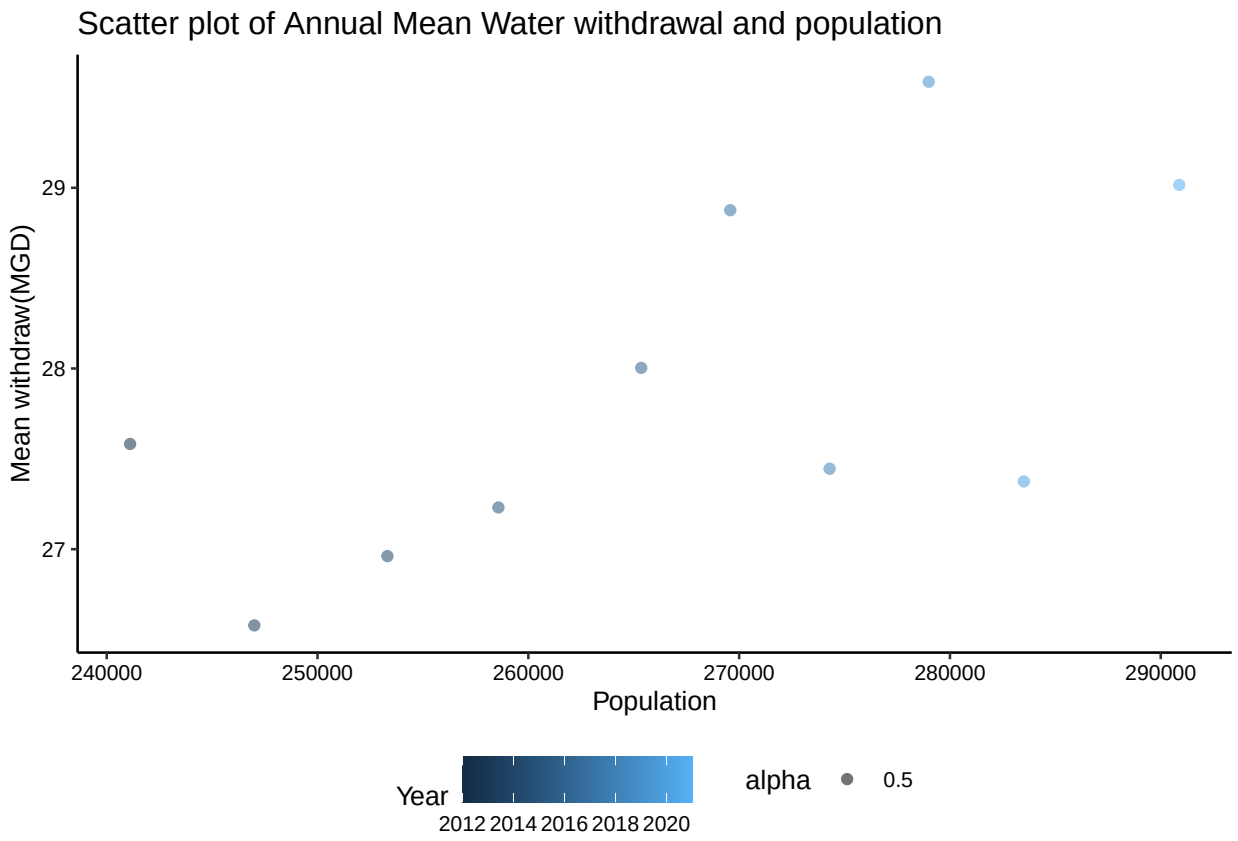
```
# Add a line and standard error for the linear regression
population_discharge_regression <- ggplot(pop_water_final, aes(x = Population,
    y = Monthly.Discharge.Wastewater)) + geom_smooth(method = "lm",
    color = "blue") + geom_point(aes(color = Year, alpha = 0.5)) +
    labs(x = "Population", y = "Monthly Wastewater Discharge(MGD)") +
    mytheme
print(population_discharge_regression)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

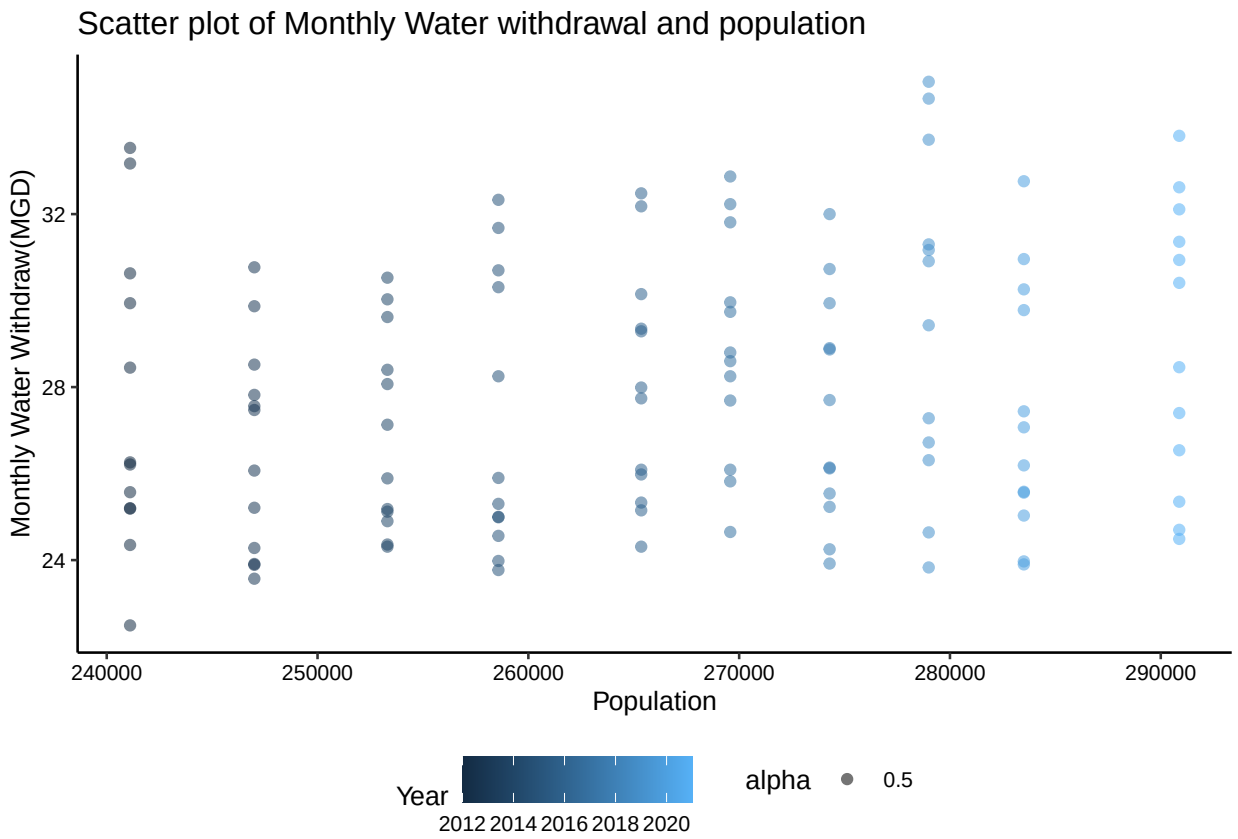


>3.3.4 Population and Wastewater withdrawal > >Similiar to population~withdrawal, this analysis uses two scatter plots:annual mean water withdrawal and population, monlth water withdrawal and population. Both that monthly water withdrawal increased as population grown.

```
# scatterplot of annual population and annual mean withdraw
population_withdrawal <- ggplot(pop_water_summaries, aes(x = Population,
    y = mean.withdraw)) + geom_point(aes(color = Year, alpha = 0.5)) +
    labs(title = "Scatter plot of Annual Mean Water withdrawal and population",
        x = "Population", y = "Mean withdraw(MGD)") + mytheme
print(population_withdrawal)
```




```
population_withdraw_scatter <- ggplot(pop_water_final, aes(x = Population,
    y = Monthly.Water.Withdraw)) + geom_point(aes(color = Year,
    alpha = 0.5)) + labs(title = "Scatter plot of Monthly Water withdrawal and population",
    x = "Population", y = "Monthly Water Withdraw(MGD)") + mytheme
print(population_withdraw_scatter)
```



>P-value(0.0334) is smaller than 0.05, showing a significant relationship between population growth and monthly water withdrawal The distribution of the residuals is fairly symmetrical, but have larger difference on the end, which means that predicted water withdrawal points fall far from the actual observed points on the extremes.Residual standard error is 15230, given that mean monthly water doscharge is 238544,the precentage error is 6.38%. All combined, over the pst decade, population is also an key influence factor on monthly water withdrawal .

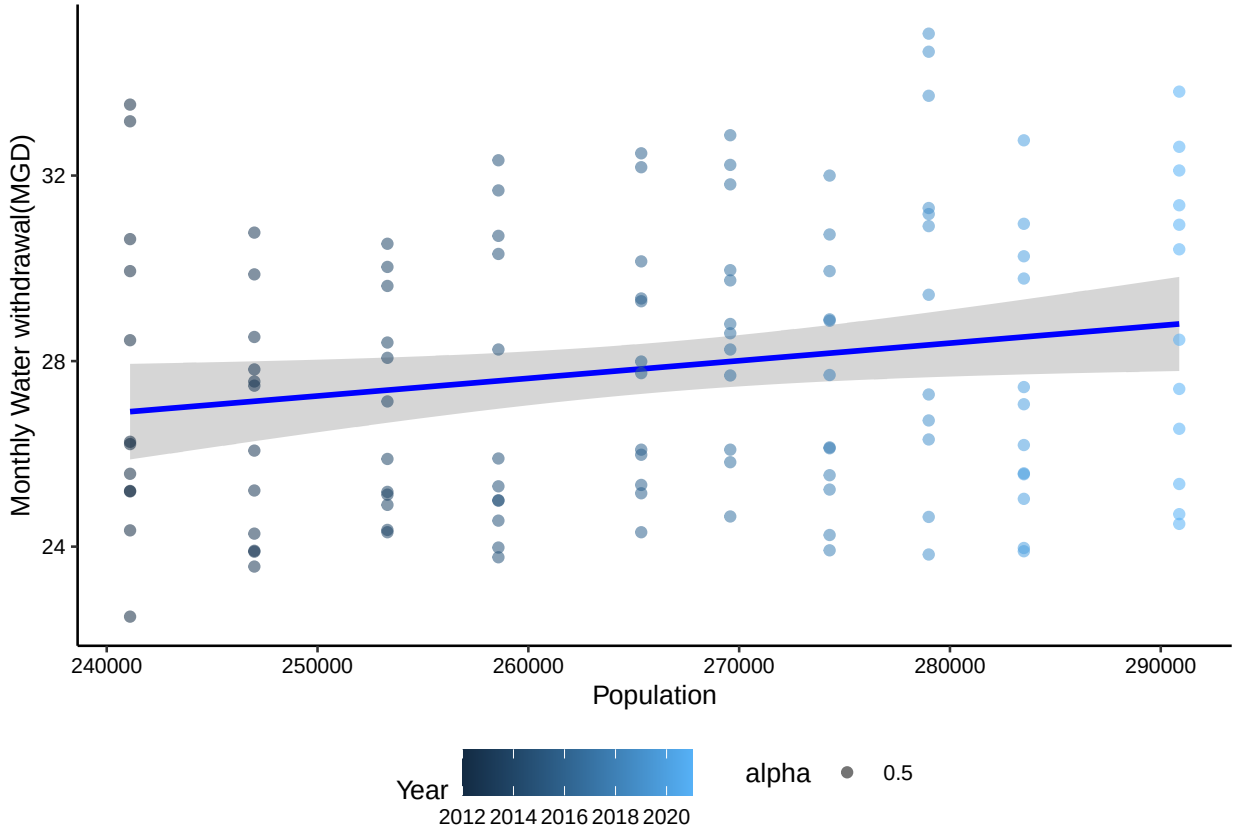
```
# Format the lm() function
population_withdraw <- lm(data = pop_water_final, Population ~
    Monthly.Water.Withdraw)
summary(population_withdraw)
```

```
##
## Call:
## lm(formula = Population ~ Monthly.Water.Withdraw, data = pop_water_final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30788 -12143   1102  11709  27974
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             238544.4    12950.7  18.419   <2e-16 ***
## Monthly.Water.Withdraw    994.6      462.1   2.153   0.0334 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 15230 on 118 degrees of freedom
## Multiple R-squared:  0.03778,    Adjusted R-squared:  0.02963
## F-statistic: 4.634 on 1 and 118 DF,  p-value: 0.03339
```

```
# Add a line and standard error for the linear regression
population_withdraw_regression <- ggplot(pop_water_final, aes(x = Population,
    y = Monthly.Water.Withdraw)) + geom_smooth(method = "lm",
    color = "blue") + geom_point(aes(color = Year, alpha = 0.5)) +
    labs(x = "Population", y = "Monthly Water withdrawal(MGD)") +
    mytheme
print(population_withdraw_regression)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



> 4. Conclusion > >Both water withdrawal and waste water discharge increased during the last 10 years, while showing different patterns within the year. Withdrawals during summer and fall are the highest when wastewater discharges are the lowest. By using linear regression, we can conclude that withdrawal,and population show correlation between each two variables, of which discharge and withdrawal have the highest significance, while popultion and monthly water discharge has highest coefficiency.