Python Applications

Tan Sein Jone

University of British Columbia

July 26, 2024

Table of contents

Table of Contents

Pollution Data

- Synthetic data on pollution levels in a province by industry and year
- Data is in a csv file
- Relatively long dataset

Table of Contents

Variable Preview

- In R studio, there's a section in the IDE that allows you to view the variables in your environment
- This is useful for checking the data types of your variables
- You can also see the first few rows of your data
- I don't know any extensions in R studio that has the same functionality as Data Wrangler in VS Code, but if you do let me know!

Table of Contents

Data Manipulation

- Data manipulation is the process of changing data to make it easier to read or more organized.
- This can involve changing the data type of a variable, removing missing values, or creating new variables.
- In R, the dplyr package is commonly used for data manipulation.

Data Manipulation

- Filter data after 2010
- Get pollution data for the transport sector
- Total pollution by province
- Applying functions

Table of Contents

Data Visualization

- Data visualization is the process of representing data graphically.
- This can help you identify patterns in the data that may not be obvious from looking at the raw data.
- In R, the ggplot2 package is commonly used for data visualization.

Data Visualization

- Create a plot of pollution levels by province
- Create a plot of pollution levels over time
- Plot pollution as a Heatmap

Table of Contents

Diff in diff

- Difference in differences is a statistical technique used to estimate the causal effect of a treatment or intervention.
- It compares the change in outcomes over time between a treatment group and a control group.
- In R, the plm package is commonly used for difference in differences analysis.

Diff in diff

- Assume there's a policy implemented in BC in 2000
- Run a diff-in-diff analysis to estimate the effect of the policy

Table of Contents

Regression Discontinuity Design (RDD)

- Regression discontinuity design is a quasi-experimental design used to estimate the causal effect of a treatment or intervention.
- It exploits the fact that individuals on either side of a cutoff point are similar in all other respects.
- In R, the rdd package is commonly used for regression discontinuity design analysis.

Regression Discontinuity Design (RDD)

- Assume there's a policy implemented in 2000
- Run a regression discontinuity design analysis to estimate the effect of the policy

Table of Contents

Tasks

- Calculate the total pollution for each province in the year 2015.
- Plot the trend of pollution over time for the 'Transport' sector, aggregating all provinces.
- Plot the total pollution over time for the top 5 provinces with the highest pollution levels in 1999.