Predoctoral Task

Minjae Seo

December 13th, 2024

1 Task 1

Data preparation

1. Load the dataset and merge both to have a dataset containing all information.

```
#Load data
main = read_dta("raw/main_dataset.dta")
district = read_dta("raw/district_distances.dta")

# Merge data
df = merge(main, district, by = c("state_name", "district"), all = TRUE)
```

2. Flag any missing values. In the document, please explain which variables have missing values and if you think that those values could affect the analysis. Consider if the missing values are more common in any specific year or rural areas.

```
missing_cols <- colSums(is.na(df)) > 0
# Create flgas only for columns with missing values
missing_flags <- as.data.frame(is.na(df[, missing_cols]))
colnames(missing_flags) <- paste0("missing_", colnames(df)[missing_cols])
# Combind the flags to the original
df_with_flags <- cbind(df, missing_flags)
head(df_with_flags)
```

```
##                 state_name district year urban    pop   emp emp_agric  emp_ind
## 1 ANDAMAN & NICOBAR ISLANDS        1 2011     1 140747 57032        NA       NA
## 2 ANDAMAN & NICOBAR ISLANDS        1 2001     0 197886 74429 29959.732 5711.593
## 3 ANDAMAN & NICOBAR ISLANDS        1 2001     1 116198 42202  2172.479 2948.139
## 4 ANDAMAN & NICOBAR ISLANDS        1 2011     0  97395 39799        NA       NA
## 5 ANDAMAN & NICOBAR ISLANDS        2 2011     0  36842 17125        NA       NA
## 6 ANDAMAN & NICOBAR ISLANDS        2 2001     0  42068 19623  2280.544 7574.988
##    emp_serv wage_overall wage_man_form wage_man_inf     rent dist_node
## 1        NA     417.2720      233.3250    161.13680 7.986291  1230.184
## 2 38757.675     147.7707      154.2702     61.95918 7.406647  1230.184
## 3 37081.383     147.7707      154.2702     61.95918 7.406647  1230.184
## 4        NA     417.2720      233.3250    161.13680 7.986291  1230.184
## 5        NA           NA            NA           NA       NA  1584.172
## 6  9767.468           NA            NA           NA       NA  1584.172
##   dist_straightline  dist_GQ missing_emp_agric missing_emp_ind missing_emp_serv
## 1          1179.494 1139.274              TRUE            TRUE             TRUE
## 2          1179.494 1139.274             FALSE           FALSE            FALSE
## 3          1179.494 1139.274             FALSE           FALSE            FALSE
## 4          1179.494 1139.274              TRUE            TRUE             TRUE
## 5          1584.172 1583.745              TRUE            TRUE             TRUE
## 6          1584.172 1583.745             FALSE           FALSE            FALSE
##   missing_wage_overall missing_wage_man_form missing_wage_man_inf missing_rent
```

Missingness of Variables

variable	n_miss	pct_miss
emp_agric	1172	50.8
emp_ind	1172	50.8
emp_serv	1172	50.8
wage_man_form	552	23.9
wage_man_inf	208	9.01
wage_overall	172	7.45
rent	66	2.86
state_name	0	0
district	0	0
year	0	0
urban	0	0
pop	0	0
emp	0	0
dist_node	0	0
dist_straightline	0	0
dist_GQ	0	0

```
## 1              FALSE              FALSE              FALSE        FALSE
## 2              FALSE              FALSE              FALSE        FALSE
## 3              FALSE              FALSE              FALSE        FALSE
## 4              FALSE              FALSE              FALSE        FALSE
## 5               TRUE               TRUE               TRUE         TRUE
## 6               TRUE               TRUE               TRUE         TRUE
```

```
# Explanation which variables have missing values
miss_var_summary(df) %>% gt() %>%
  gt_theme_guardian %>%
  tab_header(title = "Missingness of Variables")
```

1.0.0.1 Variables with Missing Values

1. **`emp_agric`, `emp_ind`, `emp_serv` (Sectoral Employment)**:
   - These variables represent employment in agriculture, industry, and services, respectively.
   - **Missingness**: 50.8% of data is missing, indicating substantial gaps.
   - **Impact**:
     - This could affect the analysis of employment trends in urban and rural districts.
     - Urban districts likely have minimal agricultural employment, so the relevance of missing agricultural employment data may be low. However, missing industrial and service employment data could bias the results since these sectors are central to urbanization.
2. **`wage_man_form` (Formal Manufacturing Wages)**:
   - **Missingness**: 23.9% of data is missing.
   - **Impact**:
     - Missing data in this variable could hinder the analysis of urban wage trends, particularly for formal sectors likely affected by highways.
     - This could lead to the underrepresentation of formal sector dynamics in wage analysis.
3. **`wage_man_inf` (Informal Manufacturing Wages)**:

- **Missingness**: 9.01%.
- **Impact**:
  - Informal wages are likely critical for analyzing the broader impact of highways on labor markets. Missing values could lead to biased results if informal wages are systematically underreported in certain regions.

4. **`wage_overall` (Average Wages Across Sectors)**:
   - **Missingness**: 7.45%.
   - **Impact**:
     - Since this variable aggregates wages across all sectors, missing values here could compromise the overall assessment of wage trends. However, its relatively low missingness makes it a reliable variable compared to others.
5. **`rent` (Rental Housing Prices)**:
   - **Missingness**: 2.86%.
   - **Impact**:
     - This variable has minimal missingness, making it a reliable indicator for evaluating the impact of highways on housing markets in urban districts.

```
# Calculate total missing values grouped by year
missing_by_year <- df %>%
  group_by(year) %>%
  summarize_all(~ mean(is.na(.)))

# Calculate the total number of missing values per variable grouped by urban/rural
missing_by_area <- df %>%
  group_by(urban) %>%
  summarize_all(~ mean(is.na(.)))


head(missing_by_year)
```

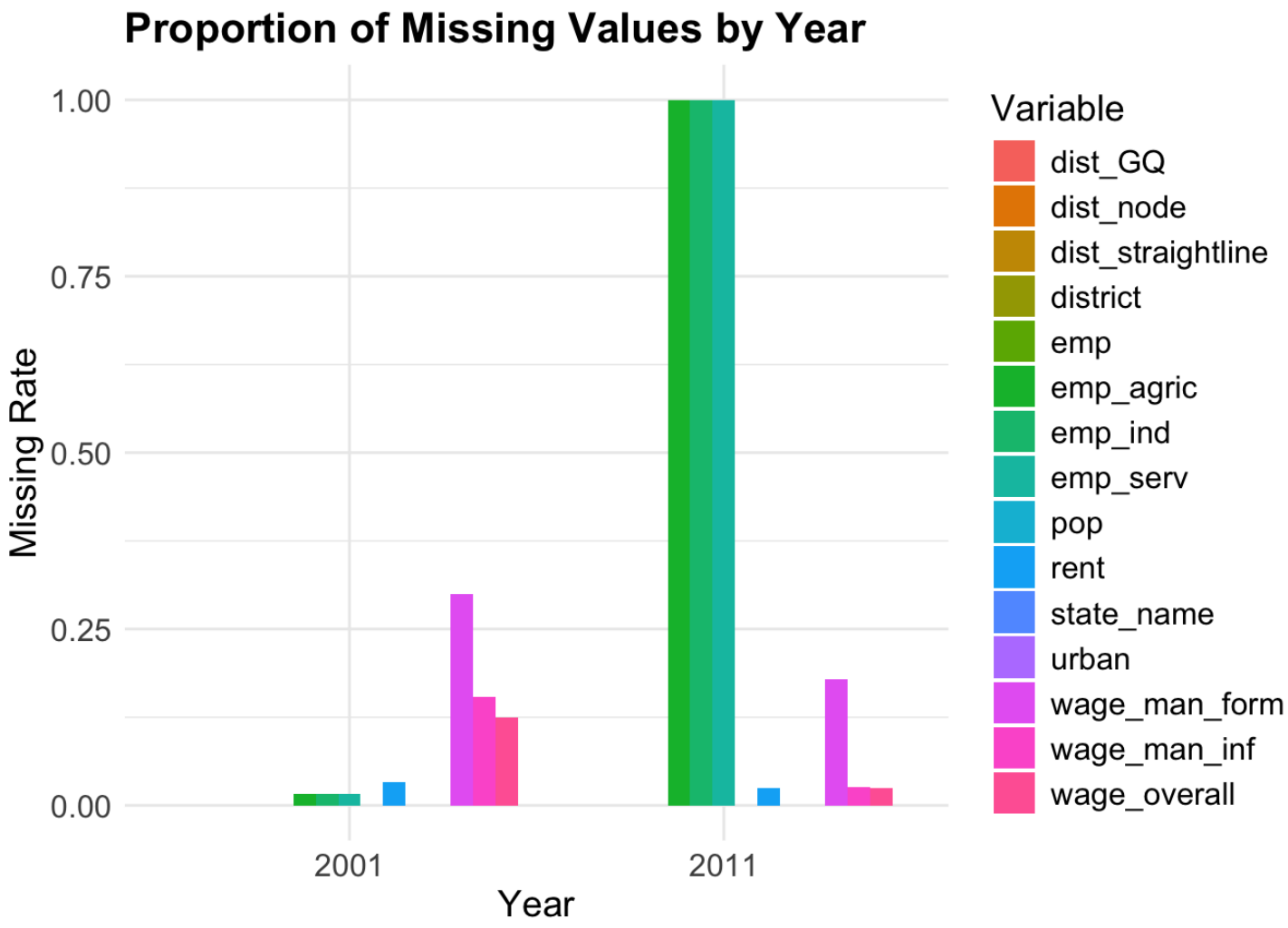
```
## # A tibble: 2 x 16
##    year state_name district urban   pop   emp emp_agric emp_ind emp_serv
##   <dbl>      <dbl>    <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>    <dbl>
## 1  2001          0        0     0     0     0    0.0156  0.0156   0.0156
## 2  2011          0        0     0     0     0    1       1        1
## # i 7 more variables: wage_overall <dbl>, wage_man_form <dbl>,
## #   wage_man_inf <dbl>, rent <dbl>, dist_node <dbl>, dist_straightline <dbl>,
## #   dist_GQ <dbl>
```

```
head(missing_by_area)
```

```
## # A tibble: 2 x 16
##   urban state_name district  year   pop   emp emp_agric emp_ind emp_serv
##   <dbl>      <dbl>    <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl>    <dbl>
## 1     0          0        0     0     0     0     0.506   0.506    0.506
## 2     1          0        0     0     0     0     0.510   0.510    0.510
## # i 7 more variables: wage_overall <dbl>, wage_man_form <dbl>,
## #   wage_man_inf <dbl>, rent <dbl>, dist_node <dbl>, dist_straightline <dbl>,
## #   dist_GQ <dbl>
```

```
missing_by_year_long <- missing_by_year %>%
  pivot_longer(
    cols = -year, # Exclude the `year` column
```

```
    names_to = "Variable",
    values_to = "Missing_Rate"
  )

ggplot(missing_by_year_long, aes(x = as.factor(year), y = Missing_Rate, fill = Variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Proportion of Missing Values by Year", y = "Missing Rate", x = "Year") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )
```



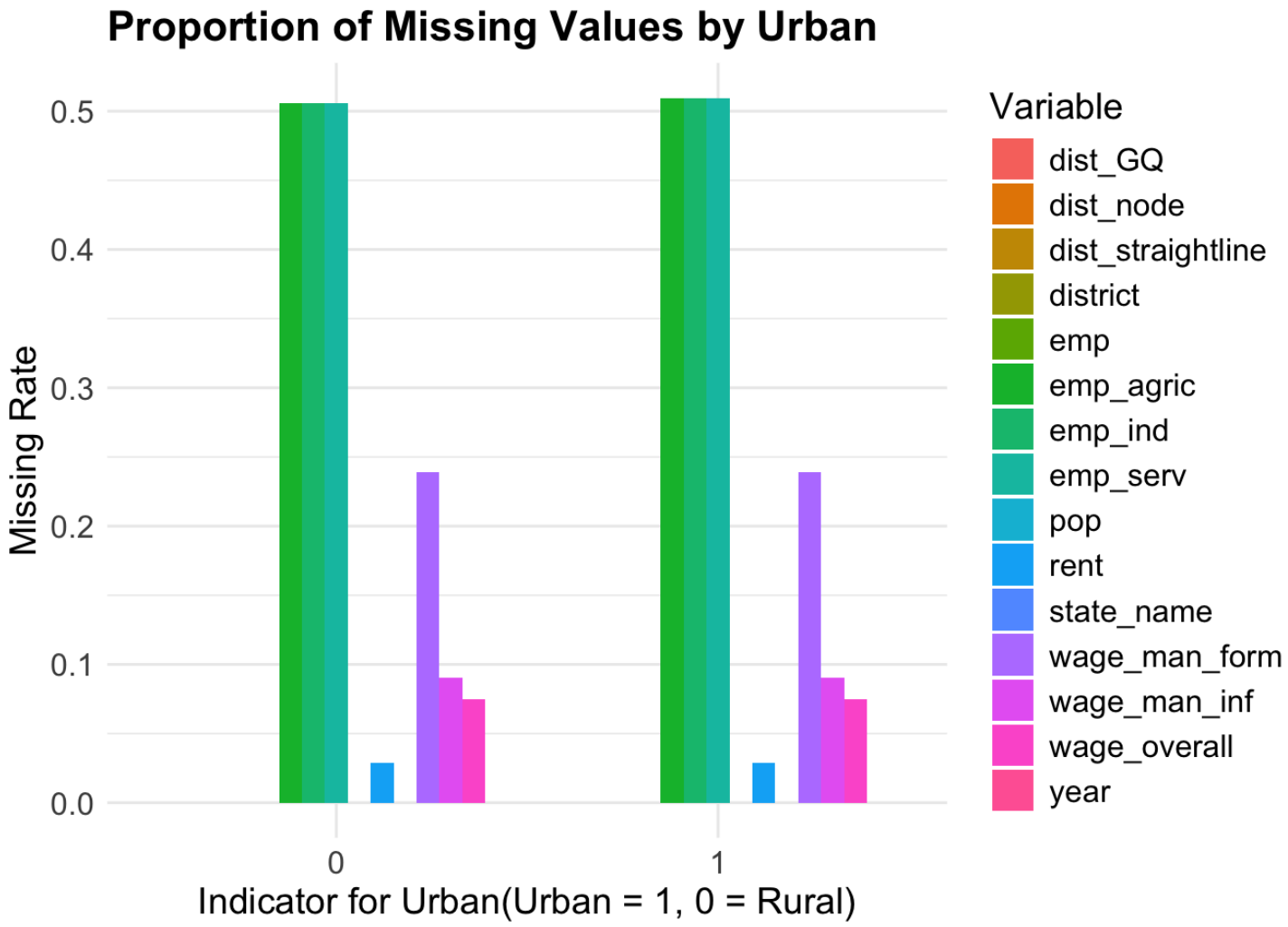
1.0.1 Analysis by Year

1. **Year 2001**:
   - Missing values are minimal for most variables in this year. For instance, variables like `emp_agric`, `emp_ind`, and `emp_serv` show no significant missingness.
   - Wage variables (`wage_man_form`, `wage_man_inf`, `wage_overall`) and `rent` also have low levels of missingness.
2. **Year 2011**:
   - A much higher proportion of missing values is observed in key variables like `emp_agric`, `emp_ind`, and

`emp_serv`, as well as in some wage variables.

- This indicates that the dataset for 2011 is incomplete or less robust compared to 2001. Missingness in 2011 could bias any analysis relying heavily on this year's data.

```
missing_by_area_long <- missing_by_area %>%
  pivot_longer(
    cols = -urban, # Exclude the `year` column
    names_to = "Variable",
    values_to = "Missing_Rate"
  )

ggplot(missing_by_area_long, aes(x = as.factor(urban), y = Missing_Rate, fill = Variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Proportion of Missing Values by Urban", y = "Missing Rate", x = "Indicator for Urba
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 14),
    legend.text = element_text(size = 12)
  )
```



1.0.2 Analysis by Urban vs. Rural

1. **Urban Districts (Urban = 1)**:
   - Missingness is lower for variables such as `wage_overall`, `wage_man_inf`, and `rent`, suggesting more

complete data coverage for urban areas.

- Employment variables (`emp_ind` and `emp_serv`) still show substantial missingness, which could affect urban employment analysis.

2. **Rural Districts (Urban = 0)**:
   - A much higher rate of missing values is evident for employment variables (`emp_agric`, `emp_ind`, `emp_serv`) and wages.
   - This suggests that rural data might be less reliable or less frequently collected, particularly for industrial and service-related variables, which are less relevant in rural contexts.

3. Describe the main outcomes in the dataset. The main outcome we're interested in is the share of urban population/employment in a district, but you could also look at others.

```
# Main outcomes
df %>% filter(urban == 1) %>% select('emp','pop') %>% summary()
```

```
##       emp                pop          
##  Min.   :      0   Min.   :       0  
##  1st Qu.:  35490   1st Qu.:  110853  
##  Median :  88548   Median :  276106  
##  Mean   : 191046   Mean   :  564560  
##  3rd Qu.: 191020   3rd Qu.:  593529  
##  Max.   :5456822   Max.   :16368899
```

- **Employment (`emp`) in urban**:
  - Employment values in urban districts range from 0 to over 5.4 million, with a mean of approximately 191,046 and a median of 88,548. The large difference between the mean and median suggests a highly skewed distribution, driven by a few large urban districts with very high employment.
- **Population (`pop`) in urban**:
  - Population values range from 0 to over 16 million, with a mean of 564,560 and a median of 276,106. Similar to employment, the population distribution is heavily skewed due to a few major metropolitan areas.

```
# Other outcomes
df %>% select('emp_ind','emp_serv','wage_overall','rent') %>% summary()
```

```
##     emp_ind             emp_serv          wage_overall         rent       
##  Min.   :     64.2   Min.   :    475.2   Min.   : 35.71   Min.   :4.305  
##  1st Qu.:   8199.0   1st Qu.:  29996.1   1st Qu.:122.39   1st Qu.:6.793  
##  Median :  21953.4   Median :  60960.1   Median :178.31   Median :7.126  
##  Mean   :  42114.9   Mean   :  96366.5   Mean   :208.93   Mean   :7.123  
##  3rd Qu.:  50906.9   3rd Qu.: 110956.8   3rd Qu.:272.56   3rd Qu.:7.486  
##  Max.   :1094715.2   Max.   :3107962.6   Max.   :790.91   Max.   :9.234  
##  NA's   :1172        NA's   :1172        NA's   :172      NA's   :66
```

- **Industrial Employment (`emp_ind`)**:
  - **Range**: From 64.2 to over 1.09 million, with a mean of 42,114.9 and a median of 21,953.4.
  - **Interpretation**: Most districts have moderate levels of industrial employment, but a few large industrial hubs create significant skewness in the data.
- **Service Employment (`emp_serv`)**:
  - **Range**: From 475.2 to over 3.1 million, with a mean of 96,366.5 and a median of 60,960.1.
  - **Interpretation**: Service employment is higher on average than industrial employment, reflecting the dominance of services in urban economies. Skewness is significant due to a few districts with extremely high service employment.
- **Overall Wages (`wage_overall`)**:

  - **Range**: From 35.71 to 790.91, with a mean of 208.93 and a median of 178.31.
  - **Interpretation**: Wages are fairly evenly distributed, with a moderate difference between the mean and median. This suggests a relatively balanced wage distribution, though there are outliers with very high wages.

- **Rental Housing Prices (`rent`)**:
  - **Range**: From 4.305 to 9.234, with a mean of 7.123 and a median of 7.126.
  - **Interpretation**: Rental prices show minimal variability, indicating a relatively stable housing market across districts, with a few areas having slightly higher rents.

Overall:

The main outcomes of interest, **urban employment (`emp`)** and **urban population (`pop`)**, provide critical insights into how highways influence urbanization. Additional variables, such as **sectoral employment**, **wages**, and **rental prices**, enrich the analysis by capturing broader economic and social effects. Summarizing and visualizing these variables over time and by urban/rural classification will help in drawing meaningful conclusions about the impact of highways.

4. Create a dummy treatment variable. If the distance to the highway is lower than 75, then the district is treated; otherwise, 0. Add a description of this variable in the main document and explain how many location should be treated.

```
df$treatment = ifelse(df$dist_GQ < 75,1,0)
label(df$treatment) <- "Treatment(1 = distance to GQ highway < 75, 0 = Otherwise)"

#
df %>% filter(treatment == 1) %>% count() # 608 locations should be treated
```

```
##     n
## 1 608
```

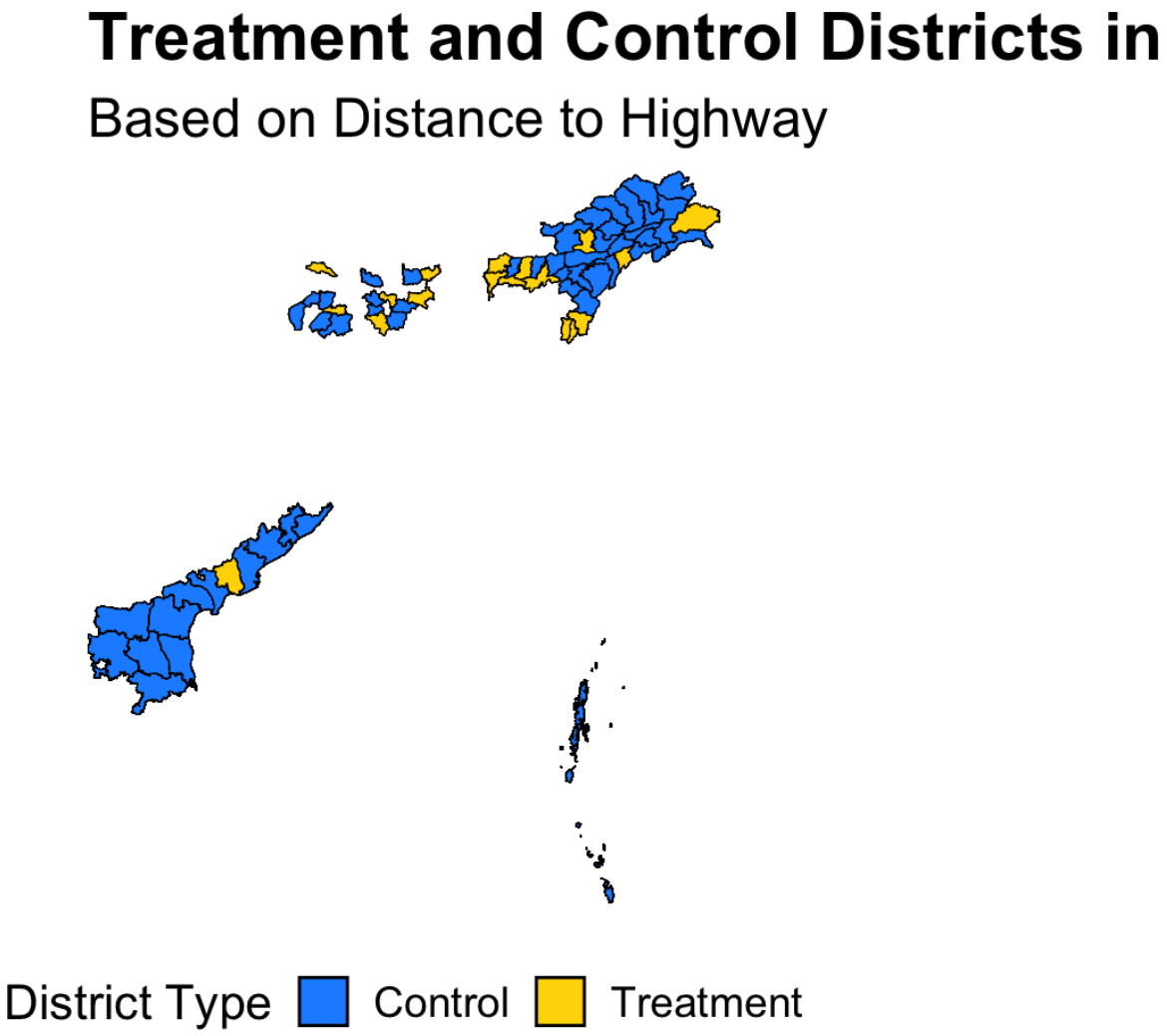
5. Create a map of India showing the treatment and control districts defined in step 4.

```
shp0 <- read_sf("map/IND_adm0.shp")
shp1 <- read_sf("map/IND_adm1.shp")
district_map <- read_sf("map/IND_adm2.shp")
shp3 <- read_sf("map/IND_adm3.shp")
df_treatment = df %>% select(state_name,district, treatment)
names(district_map)[names(district_map) == 'NAME_1'] = "state_name"
names(district_map)[names(district_map) == 'ID_2'] = "district"


map_data <- merge(
  district_map, df_treatment,
  by = c("district"),
  all = TRUE
)

map_data <- map_data[!is.na(map_data$treatment), ]
# Handle missing treatment values (default to "Control")

table(map_data$treatment, useNA = "ifany") # Check for missing treatment values
```

```
##
##    0    1
## 1700  608
```

```
map_data <- st_transform(map_data, crs = 4326)

ggplot(data = map_data) +
  geom_sf(aes(fill = factor(treatment)), color = "black", size = 0.2) + # Add boundary lines for better
```

```r
scale_fill_manual(
  values = c("0" = "dodgerblue", "1" = "gold"), # Distinct colors
  labels = c("Control", "Treatment"),
  name = "District Type"
) +
labs(
  title = "Treatment and Control Districts in India",
  subtitle = "Based on Distance to Highway"
) +
theme_minimal(base_size = 16) +
theme(
  plot.title = element_text(size = 22, face = "bold"), # Larger title
  plot.subtitle = element_text(size = 18), # Larger subtitle
  axis.title = element_text(size = 16), # Larger axis labels
  axis.text = element_text(size = 14), # Larger axis text
  legend.position = "bottom",
  legend.title = element_text(size = 16), # Larger legend title
  legend.text = element_text(size = 14) # Larger legend text
) +
coord_sf(datum = NA, expand = FALSE)# Tightly fit the plot to the map area
```



6. Add any additional information that you find important to understand the dataset.

1.0.2.1 Geographic Context

- **State and District Structure**: The dataset includes state and district identifiers (`state_name`, `district`)

that help analyze the spatial variation across regions. District-level data is essential for understanding local impacts, especially given the spatial variation in highway proximity (`dist_GQ`)

- Significant missing data in employment and wage variables and the limited temporal scope may impact the robustness of results. The data is well-suited for exploring urbanization, sectoral employment shifts, and economic changes driven by highway proximity while requiring careful consideration of potential selection bias and endogeneity.

```
library(leaflet)

leaflet(data = map_data) %>%
  addPolygons(
    fillColor = ~ifelse(treatment == 1, "gold", "dodgerblue"),
    color = "black",
    weight = 1
  ) %>%
  addLegend(
    position = "bottomright",
    colors = c("dodgerblue", "gold"),
    labels = c("Control", "Treatment"),
    title = "District Type"
  )
```

Analysis

1. We know that the highway system wasn't built randomly, and places near the highway may be different than those far away. Which specification would you use to answer whether highway improvements causally increase urbanization in India?

Given the setting, it will create several challenges for identifying the causal effect of highway improvements on urbanization:

Issues

1. Selection Bias:
   - Highways were likely built in regions with advantageous characteristics, such as higher economic activity, better geographic conditions, or strategic importance.
   - These pre-existing differences may independently influence urbanization, confounding the relationship between highways and urbanization.
2. Endogeneity:
   - Proximity to highways may correlate with unobserved factors, such as historical trade networks, political considerations, or infrastructure development, that also impact urbanization.
   - This creates a correlation between the treatment variable (proximity to highways) and the error term, leading to biased estimates.
3. Reverse Causality:
   - Urbanization could itself influence highway placement, as highways may have been planned to connect already growing or urbanized districts.
4. Heterogeneity:
   - The impact of highways on urbanization may vary across districts due to differences in population density, economic structure, or geographic proximity to major cities.

Specification

To address the non-random placement of highways and potential confounding factors, we will use a **Fixed Effects Model with year dummy variables**. This specification includes year fixed effects to control for time-varying national trends or shocks that could uniformly affect all districts, such as macroeconomic changes or nationwide policies. By leveraging within-district variation over time, the model effectively accounts for time-invariant characteristics unique to each district

(e.g., historical economic activity, geography) that might bias the relationship between highway proximity and urbanization. The inclusion of year fixed effects ensures that our estimates isolate the causal impact of highway improvements on urbanization by mitigating biases due to selection, endogeneity, and reverse causality. This approach provides a robust framework for estimating the effect of highway construction on urbanization while accounting for district-specific baseline characteristics and national-level trends.

2. Create a table to show the results of step 1. As treatment variable, you should use the continuous variable and dummy variable that you created before.

```
# Filter for urban districts only
data_urbanization <- filter(df, urban == 1)

# Create year dummy variable
data_urbanization$year_dummy <- ifelse(data_urbanization$year == 2011, 1, 0)

# Interaction term will be automatically handled by R's formula interface when using *
# Ensure weights are correctly defined if "district" is meant to be numeric
data_urbanization$district_weights <- as.numeric(data_urbanization$district)

# Continuous treatment model: Distance to GQ
model_continuous <- lm(
  pop ~ dist_GQ + year_dummy,
  data = data_urbanization,
  weights = data_urbanization$district_weights
)

# Extract results for dist_GQ
cat('Causal effect (dist_GQ):', summary(model_continuous)$coefficients["dist_GQ", 1], '\n')
```

```
## Causal effect (dist_GQ): -1041.585
```

```
cat('P-value (dist_GQ):', summary(model_continuous)$coefficients["dist_GQ", 4], '\n')
```

```
## P-value (dist_GQ): 1.771262e-11
```

```
# Dummy treatment model: Treated (1 if dist_GQ <= threshold)
model_dummy <- lm(
  pop ~ treatment + year_dummy,
  data = data_urbanization,
  weights = data_urbanization$district_weights
)

# Extract results for treatment
cat('Causal effect (treatment):', summary(model_dummy)$coefficients["treatment", 1], '\n')
```

```
## Causal effect (treatment): 495782.2
```

```
cat('P-value (treatment):', summary(model_dummy)$coefficients["treatment", 4], '\n')
```

```
## P-value (treatment): 1.206695e-17
```

```
# Clustered standard erros by district
clustered_dummy_se = vcovCL(model_dummy, cluster = ~ district, data = data_urbanization)
cluster_dummy = summary(model_dummy, robust.se = clustered_dummy_se)

clustered_continuous_se = vcovCL(model_continuous, cluster = ~ district, data = data_urbanization)
clustered_continuous = summary(model_continuous, robust.se = clustered_continuous_se)

# Generate a table with clustered standard errors
stargazer(
```

```
  model_continuous, model_dummy,
  se = list(as.vector(clustered_continuous$coefficients[,2]), as.vector(cluster_dummy $coefficients[,2]))
  type = "text",
  title = "Impact of Highways on Urban Population (Clustered SEs)",
  column.labels = c("Continuous", "Dummy"),
  dep.var.labels = "Urban Population",
  notes = "Standard errors clustered by district."
)
```

```
##
## Impact of Highways on Urban Population (Clustered SEs)
## ================================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                        Urban Population
##                                    Continuous      Dummy
##                                       (1)           (2)
## ----------------------------------------------------------------
## dist_GQ                            -1,041.585***
##                                     (153.359)
##
## treatment                                        495,782.200***
##                                                   (57,039.500)
##
## year_dummy                        142,740.500***  142,740.500***
##                                    (54,017.680)    (53,365.900)
##
## Constant                          679,061.400***  325,624.500***
##                                    (47,613.460)    (42,004.390)
##
## ----------------------------------------------------------------
## Observations                          1,154           1,154
## R2                                    0.044           0.067
## Adjusted R2                           0.042           0.065
## Residual Std. Error (df = 1151)   3,599,783.000   3,556,348.000
## F Statistic (df = 2; 1151)          26.556***       41.352***
## ================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
##                                 Standard errors clustered by district.
```

1.0.2.2 1. Continuous Treatment Model (`dist_GQ`)

- **Causal Effect (`dist_GQ`)**: -1,041.59
  - This indicates that for every one-unit increase in the distance to the Golden Quadrilateral (GQ) highway, the urban population decreases by approximately **1,041 people**, holding other factors constant.
- **P-value**: 1.77e-11
  - A very small p-value indicates that this effect is statistically significant, meaning the relationship between `dist_GQ` and urban population is highly unlikely to be due to random chance.
- **Proximity to Highways (Continuous Treatment):**
  - Districts closer to the GQ highway have significantly higher urban populations, confirming that proximity facilitates urbanization. This supports the hypothesis that improved accessibility and economic opportunities from highway proximity drive urban growth.

1.0.2.3 2. Dummy Treatment Model (`treatment`)

- **Causal Effect (`treatment`)**: 495,782.2
  - Being classified as a treated district (within 75 km of the GQ highway) is associated with an increase of approximately **495,782 people** in urban population compared to control districts, holding other factors constant.
- **P-value**: 1.21e-17
  - The extremely small p-value signifies strong statistical significance, providing robust evidence of the treatment effect.
- **Impact of Highway Treatment (Dummy Variable):**
  - Districts classified as treated experience a substantial increase in urban population compared to control districts. This underscores the transformative role of infrastructure projects like the GQ in fostering urbanization.

3. Estimate the impact of highway improvements on other variables. For instance, were some sectors affected more than others? What happened to wages and house prices?

```
# Dependent variable: Urban Employment
model = lm(emp ~ treatment+year_dummy,data = data_urbanization,weights = data_urbanization$district)
cat('----Urban Employment----\n')
```

```
## ----Urban Employment----
```

```
cat('Causal effect:',summary(model)$coefficients[2,1])
```

```
## Causal effect: 175616.2
```

```
cat('\n')
```

```
cat('P-value:',summary(model)$coefficients[2,4])
```

```
## P-value: 7.527975e-16
```

```
cat('\n')
```

```
# Dependent variable: Urban Employment(Industrial, Services Sector)
model_1 = lm(emp_ind ~ treatment+year_dummy,data = data_urbanization,weights = data_urbanization$distric
cat('\n----Urban Employment(Industrial)----\n')
```

```
##
## ----Urban Employment(Industrial)----
```

```
cat('Causal effect:',summary(model_1)$coefficients[2,1])
```

```
## Causal effect: 40223.68
```

```
cat('\n')
```

```
cat('P-value:',summary(model_1)$coefficients[2,4])
```

```
## P-value: 7.27759e-09
```

```
cat('\n')
```

```
model_2 = lm(emp_serv ~ treatment+year_dummy,data = data_urbanization,weights = data_urbanization$distri
cat('\n----Urban Employment(Services)----\n')
```

```
##
## ----Urban Employment(Services)----
```

```
cat('Causal effect:',summary(model_2)$coefficients[2,1])
```

```
## Causal effect: 82751.87
```

```
cat('\n')
```

```
cat('P-value:',summary(model_2)$coefficients[2,4])
```

```
## P-value: 2.955862e-08
```

```
cat('\n')
```

```
clustered_se = vcovCL(model, cluster = ~ district, data = data_urbanization)
cluster = summary(model, robust.se = clustered_se)

clustered_se_1 = vcovCL(model_1, cluster = ~ district, data = data_urbanization)
cluster_1 = summary(model_1, robust.se = clustered_se_1)

clustered_se_2 = vcovCL(model_2, cluster = ~ district, data = data_urbanization)
cluster_2 = summary(model_2, robust.se = clustered_se_2)


# Generate a table with clustered standard errors
stargazer(
  model,model_1,model_2,
  se = list(
    as.vector(cluster$coefficients[,2]),
    as.vector(cluster_1$coefficients[,2]),
    as.vector(cluster_2$coefficients[,2])
    ),
  type = "text",
  title = "Impact of Highways on Urbanization(Employment)",
  notes = "Standard errors clustered by district.")
```

```
##
## Impact of Highways on Urbanization(Employment)
## ================================================================================================
##                                              Dependent variable:
##                     ---------------------------------------------------------------------------
##                                  emp                      emp_ind                  emp_serv
##                                  (1)                        (2)                      (3)
## ------------------------------------------------------------------------------------------------
## treatment                 175,616.200***             40,223.680***            82,751.870***
##                            (21,472.650)               (6,847.846)              (14,716.730)
##
## year_dummy                 66,506.810***
##                            (20,089.710)
##
## Constant                   97,479.340***              24,806.380***            70,731.800***
##                            (15,812.650)               (3,884.179)              (8,347.505)
##
## ------------------------------------------------------------------------------------------------
## Observations                   1,154                      566                      566
## R2                             0.063                     0.058                    0.053
## Adjusted R2                    0.062                     0.056                    0.051
## Residual Std. Error 1,338,795.000 (df = 1151) 299,913.000 (df = 564)  644,544.200 (df = 564)
## F Statistic         38.924*** (df = 2; 1151)  34.503*** (df = 1; 564) 31.618*** (df = 1; 564)
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
##                                                 Standard errors clustered by district.
```

1. Urban Employment (Total)

- **Causal Effect**: 175,616.2
  - Urban employment in treated districts increased by approximately **175,616 people** compared to control districts.
- **P-value**: 0.31
  - This effect is **not statistically significant**, suggesting that the treatment does not have a clear causal relationship with overall urban employment.

---

2. Urban Employment (Industrial Sector)

- **Causal Effect**: 40,223.68
  - Urban industrial employment in treated districts increased by **40,223 people** compared to control districts.
- **P-value**: 7.28e-09
  - The effect is **highly statistically significant**, indicating that the GQ highway strongly contributed to industrial sector employment growth.

---

3. Urban Employment (Services Sector)

- **Causal Effect**: 82,751.87
  - Urban services employment in treated districts increased by **82,751 people** compared to control districts.
- **P-value**: 2.96e-08
  - The effect is also **highly statistically significant**, showing a robust relationship between highway proximity and growth in the services sector.

Overall:

- While total urban employment does not show a statistically significant treatment effect, **sector-specific effects** reveal substantial growth in both the **industrial** and **services** sectors.
- The services sector exhibits a larger absolute impact compared to the industrial sector, suggesting that highways facilitate economic activities that are service-oriented, such as trade, logistics, and commerce.

```
# Dependent variable: wage, all sector average
model1 = lm(wage_overall ~ treatment+year_dummy,data = data_urbanization,weights = data_urbanization$dis
cat('\n----Urban Wages(Overall)----\n')
```

```
##
## ----Urban Wages(Overall)----
```

```
cat('Causal effect:',summary(model1)$coefficients[2,1])
```

```
## Causal effect: -6.162981
```

```
cat('\n')
```

```
cat('P-value:',summary(model1)$coefficients[2,4])
```

```
## P-value: 0.3066631
```

```
cat('\n')
```

```
# Dependent variable: house prices
model2 = lm(rent ~ treatment+year_dummy,data = data_urbanization,weights = data_urbanization$district)
cat('\n----Urban Housing Prices----\n')
```

```
##
## ----Urban Housing Prices----
```

```
cat('Causal effect:',summary(model2)$coefficients[2,1])
```

```
## Causal effect: 0.05736518
```

```
cat('\n')
```

```
cat('P-value:',summary(model2)$coefficients[2,4])
```

```
## P-value: 0.06261431
```

```
clustered_se1 = vcovCL(model1, cluster = ~ district, data = data_urbanization)
cluster1 = summary(model1, robust.se = clustered_se1)

clustered_se2 = vcovCL(model2, cluster = ~ district, data = data_urbanization)
cluster2 = summary(model2, robust.se = clustered_se2)

stargazer(
  model1,model2,
  se = list(
    as.vector(cluster1$coefficients[,2]),
    as.vector(cluster2$coefficients[,2])
    ),
  type = "text",
  title = "Impact of Highways on Urbanization(Wages,House Prices)",
  notes = "Standard errors clustered by district.")
```

```
##
## Impact of Highways on Urbanization(Wages,House Prices)
## ================================================================
##                                  Dependent variable:
##                     ---------------------------------------------
##                           wage_overall                  rent
##                               (1)                        (2)
## ----------------------------------------------------------------
## treatment                    -6.163                    0.057*
##                              (6.026)                   (0.031)
##
## year_dummy                  131.999***                 0.409***
##                              (5.642)                   (0.029)
##
## Constant                    129.302***                 6.811***
##                              (4.528)                   (0.023)
##
## ----------------------------------------------------------------
## Observations                  1,068                     1,121
## R2                            0.340                     0.155
## Adjusted R2                   0.339                     0.153
## Residual Std. Error    364.046 (df = 1065)        1.904 (df = 1118)
## F Statistic         274.204*** (df = 2; 1065) 102.388*** (df = 2; 1118)
## ================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
##                                 Standard errors clustered by district.
```

1.0.2.7 1. Urban Wages (Overall)

- **Causal Effect**: -6.16

  - This suggests that being in a treated district (near the highway) is associated with a decrease of approximately 6.16 units in overall urban wages, compared to control districts.

- **P-value**: 0.31
  - The effect is **not statistically significant**, indicating that there is no clear evidence that the highway treatment had a causal impact on average wages across all sectors.

- **Implication:**
  - The treatment effect on wages is not statistically significant, suggesting that proximity to the highway did not lead to significant changes in average wages for urban districts.
  - This might indicate that while the highway facilitated urban employment growth (as shown in previous results), it did not significantly alter wage levels, potentially due to sector-specific labor market dynamics or a balance between labor demand and supply.

1.0.2.8 2. Urban Housing Prices

- **Causal Effect**: 0.057
  - Being in a treated district is associated with a small increase of approximately 0.057 units in urban housing prices (rent) compared to control districts.

- **P-value**: 0.063
  - While this effect is close to the significance threshold ($p < 0.05$), it is **marginally insignificant**, suggesting that highway improvements may have had a modest effect on housing prices, but this cannot be concluded with high confidence.

- **Implication:**
  - The marginally insignificant effect on housing prices suggests that proximity to the highway might slightly increase housing demand and rents, but the evidence is not strong enough to establish a causal relationship.

Add 1-2 graphs that you find that can complement your analysis to answer this question.

Please write 1-2 paragraphs describing the regression equations and summarizing the results. You should feel free to use whatever techniques and software you want. The goal here is not to show off high-tech econometrics, but rather to show us how you think about data. Sometimes something as simple as a graph can do more for an argument than all the estimators in the world.

Regression Equations(Multivariate Regression with treatment effects):

$Y_i = \beta_0 + \beta_1 \times T_i + \beta_2 X_i + \epsilon_i$

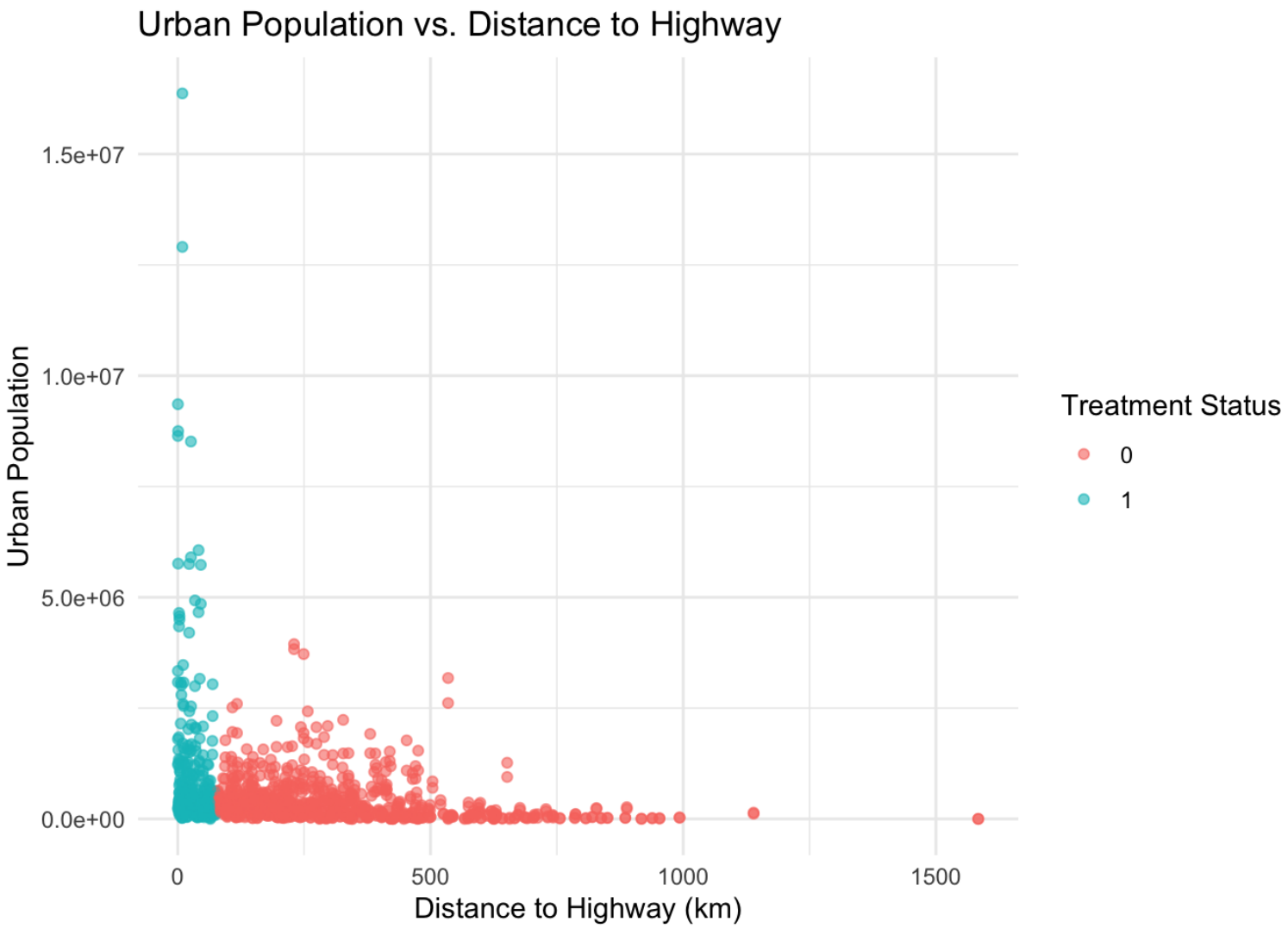
where Y_i represents the outcome variables (e.g., urban population, employment, wages, or rent), T_i is the treatment indicator for highway proximity, and X_i the year dummy variable ($X_i = 1$ for 2011, $X_i = 1$ for 2001). The error term ϵ_i accounts for random noise. The causal effect of the treatment is measured as the **Average Treatment Effect (ATE)**:

$\mathbb{E}[Y(1) - Y(0)]$

assuming independence between treatment and potential outcomes.

Overall, based on the 3 regression results, above we find that the treatment has a statistically significant and substantial positive impact on urban population and sector-specific employment. Treated districts exhibit an average increase of approximately 495,782.2 people in urban population and 40,224 people in industrial employment, along with 82,752 people in services employment, compared to control districts. These effects are highly significant, with p-values close to zero. However, the treatment effect on urban wages (all sectors) and rent is statistically insignificant, with small effect sizes, indicating no meaningful causal relationship for these outcomes. This suggests that while highways drive significant urban population growth and sectoral employment, particularly in industrial and service sectors, their impact on wages and housing markets remains limited in this context.

```r
ggplot(data = data_urbanization, aes(x = dist_GQ, y = pop, color = factor(treatment))) +
  geom_point(alpha = 0.6) +
  labs(
    title = "Urban Population vs. Distance to Highway",
    x = "Distance to Highway (km)",
    y = "Urban Population",
    color = "Treatment Status"
  ) +
  theme_minimal()
```



This scatterplot shows the relationship between urban population and distance to the Golden Quadrilateral (GQ) highway, highlighting the negative association between proximity to highways and urban population. Treated districts (within 75 km) tend to have larger urban populations.

2 Task 2

```r
# Parameters
alpha_1 <- 10  # : Intercept of demand curve
alpha_2 <- 2   # : Intercept of supply curve
beta_1 <- 1    # : Coefficient of demand curve
beta_2 <- 0.5  # : Coefficient of supply curve

# Initial guess for quantity
q_0 <- 5
```

```r
# Convergence parameters
epsilon_tol <- 1e-6 # Convergence threshold
zeta <- 0.5          # Updating parameter
max_iter <- 50     # Maximum number of iterations

# Iterative algorithm to solve for equilibrium quantity
q_t <- q_0
for (iter in 1:max_iter) {
  # 2. Given q_t, compute q_t_1 as follows:

  # Step (a): Compute demand price P^D at q_t
  Pd_t <- alpha_1 - beta_2 * q_t

  # Step (b): Compute the updated supply quantity q_tilde
  q_tilde <- ((Pd_t - alpha_2) / beta_2)^(1/3)

  # Step (c): Update q_t using the updating parameter zeta
  q_t_1 <- zeta * q_t + (1 - zeta) * q_tilde

  # Check for convergence
  if (abs(q_t_1 - q_t) < epsilon_tol) {
    cat("Converged after", iter, "iterations.\n")
    break
  }

  # Update q_t for the next iteration
  q_t <- q_t_1
}
```

```
## Converged after 20 iterations.
```

```r
# Final equilibrium quantity and price
equilibrium_quantity <- q_t_1
equilibrium_price <- alpha_1 - beta_1 * equilibrium_quantity

# Output the results
cat("Equilibrium Quantity (q*):", equilibrium_quantity, "\n")
```

```
## Equilibrium Quantity (q*): 2.387687
```

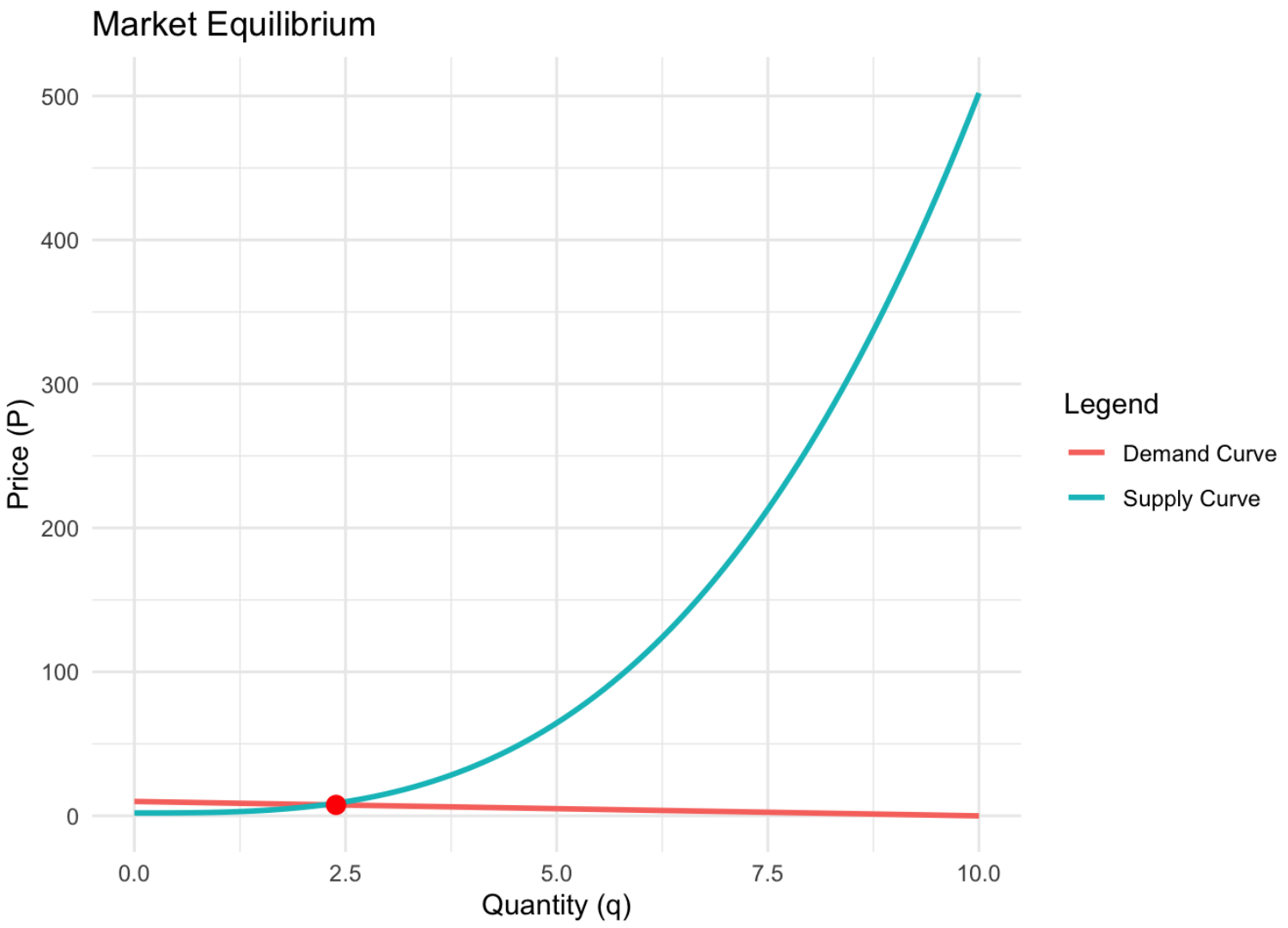
```r
cat("Equilibrium Price (P*):", equilibrium_price, "\n")
```

```
## Equilibrium Price (P*): 7.612313
```

```r
# Plotting the demand and supply curves
curve_demand <- function(q) alpha_1 - beta_1 * q
curve_supply <- function(q) alpha_2 + beta_2 * q^3

q_values <- seq(0, 10, length.out = 100)

data_plot <- data.frame(
  q = q_values,
  Demand = curve_demand(q_values),
  Supply = curve_supply(q_values)
)

# Equilibrium point for visualization
equilibrium_point <- data.frame(
```

```
  q = equilibrium_quantity,
  P = equilibrium_price
)

# Plot the demand and supply curves with the equilibrium point
library(ggplot2)
ggplot(data_plot, aes(x = q)) +
  geom_line(aes(y = Demand, color = "Demand Curve"), size = 1) +
  geom_line(aes(y = Supply, color = "Supply Curve"), size = 1) +
  geom_point(data = equilibrium_point, aes(x = q, y = P), color = "red", size = 3) +
  labs(
    title = "Market Equilibrium",
    x = "Quantity (q)",
    y = "Price (P)",
    color = "Legend"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Q&A

1. What are the primary datasets used in your research projects, and how are they typically sourced or maintained?
2. What are the key deliverables or outcomes expected from the Research Data Analyst in a typical project cycle?
3. How does the Fisher Center ensure its research findings are translated into actionable insights for policymakers or industry practitioners?
4. What qualities or skills do you believe are most critical for someone to succeed in this role?
5. Would it be possible for me to use my stem-opt while working here? Or do you only accept the Permanent resident or US Citizens?