PS 531: Final Project – Pre-Analysis Plan

Akayla A. Henson

13 December 2024

Pre-Analysis Plan: Racial Profiling in Police Stop Outcomes

Introduction: Racial Disparities in Police Stop Outcomes (Question 1)

Traffic stops are one of the most common ways police officers interact with the public. However, during these routine encounters, those who identify as people of color often report experiencing racial profiling, which remains a saalien social issue in the United States (Grogger and Ridgeway 2006). In the past decade, highly publicized fatal incidents involving police and people of color have drawn widespread attention (Nadal et al. 2017). While Black Americans have long been disproportionately targeted for harassment and violence by police, these recent events and subsequent studies have amplified public scrutiny of law enforcement. This heightened awareness has influenced social media discourse, mainstream media narratives, community activism, educational discussions, and even political campaigns. To better understand public perceptions of the police, examining their interactions during traffic stops offers a valuable starting point. The substantive question in this study is: Are individuals of certain racial or ethnic groups more likely to experience adverse outcomes, such as searches or frisks, after police stops in Louisville, KY? This question is important to help understand potential racial disparities in law enforcement practices, which have implications for civil rights, public trust in police, and broader societal equity.

The theory underlying this analysis is rooted in Critical Race Theory (CRT), which posits that structural racism is embedded in institutions like law enforcement, leading to disparate treatment of racial and ethnic minorities (Bridges 2019). CRT suggests that these disparities are not merely the result of individual bias but are also the outcome of systemic factors that shape policing policies and practices. This study will test the hypothesis that racial disparities in stop outcomes exist. Understanding whether racial disparities exist in police stop outcomes would aid in addressing systemic inequities in law enforcement. Disparities in stop outcomes can undermine public trust in police, particularly among minority communities, leading to strained relationships and decreased cooperation. This issue has become especially salient in the wake of nationwide protests against police violence and racial injustice, such as those following the death of Breonna Taylor in Louisville. Additionally, this study could inform policymakers and law enforcement agencies about the need for reforms to reduce disparities and improve accountability. By providing evidence of whether and where disparities exist, this research can contribute to efforts aimed at creating fairer and more equitable policing practices, which are essential for fostering community safety and trust.

Hypothesis (Question 2)

This study will assess one hypothesis that arises from CRT: Individuals from racial or ethnic minority groups are more likely to experience adverse outcomes (i.e., searches or frisks) during stops compared to White individuals. This hypothesis is grounded in CRT's assertion that structural racism within institutions like law enforcement results in differential treatment of racial and ethnic minorities. CRT posits that these disparities are not just due to individual prejudice but also to institutional policies and practices that disproportionately impact marginalized groups. CRT provides a framework for understanding how race and power dynamics intersect to shape societal structures, including policing. According to CRT, law enforcement practices often reflect broader societal biases, leading to unequal treatment based on race or ethnicity. This theory justifies the expectation that racial disparities in police stop outcomes will exist, even after accounting for other factors like time and locatioin. The hypothesis is supported by previous empirical research showing that Black and Hispanic individuals are disproportionately stopped, searched, and arrested compared to White individuals (Grogger and Ridgeway 2006; Nadal et. al. 2017).

Data and Research Design (Question 3)

This study investigates racial disparities in police stop outcomes in Louisville, KY, using data from The Stanford Open Policing Project. The dataset covers police stops conducted between June 2014 and December 2020, providing information on the race and ethnicity of individuals stopped, the location and time of the stop, and outcomes such as searches, citations, warnings, or frisks. This dataset is well-suited for analyzing racial disparities, as it captures individual-level data on police interactions and includes contextual factors like location of the stop. This contextual variable allows for an analysis that controls for situational differences, enabling the isolation of the effect of race or ethnicity on stop outcomes.

This study employs a quasi-experimental design using propensity score matching (PSM) to address the observational nature of the data and ensure interpretable comparisons. By matching individuals based on observable covariates like location and time of stop, the design minimizes confounding and creates an as-if-randomized comparison. This approach aligns with the insights from Paul Rosenbaum's *Design of Observational Studies* (2010), which emphasizes balancing covariates to strengthen causal claims in non-experimental data. The matched data will be analyzed using descriptive statistics, balance tests, and regression models. The primary analysis compares the likelihood of adverse outcomes between racial and ethnic groups after controlling for contextual factors through matching. Balance diagnostics will confirm whether the matching procedure has successfully created comparable groups, while regression models will estimate the causal effect. To ensure robustness, the study will conduct sensitivity analyses to evaluate the findings and assess potential biases or the influence of unmeasured confounders.

Advantages and Disadvantages (Question 4)

The observational causal design employed in this study offers several advantages for addressing the substantive question of racial disparities in police stop outcomes. First, it uses real-world data from the Stanford Open Policing Project, which provides detailed information on police stops, allowing for an analysis of individual-level interactions. The inclusion of contextual variables such as location enables us to account for situational factors that might otherwise confound the relationship between race and stop outcomes. By using propensity score matching (PSM), the study approximates a counterfactual comparison, which is needed for causal inference in an observational setting. This matching process helps to reduce bias from observable confounders, ensuring that comparisons are made between similar stops, thereby increasing the credibility of the findings. Additionally, the large sample size and detail in the dataset support subgroup analyses, allowing for exploration of disparities across different racial and ethnic groups.

However, this research design has limitations that must be acknowledged. A primary disadvantage is its reliance on observational data, which inherently precludes randomization and leaves the study vulnerable to unobserved confounding (Morgan and Winship 2015). While propensity score matching addresses observed differences between groups, it cannot account for unmeasured variables, such as broader systemic factors influencing stop outcomes. Another challenge is the potential for measurement error in the dataset, as data quality may vary across different reporting officers or departments. Also, matching techniques can result in a loss of data if many cases do not find suitable matches, potentially reducing statistical power and generalizability. Finally, while the design approximates causal inference, it cannot definitively establish causation, limiting the strength of the conclusions drawn from the analysis.

Measures (Question 5)

The dataset includes several variables needed for analyzing racial disparities in police stop outcomes. Key measures include the date and time of each stop, which provide time context, and division and beat, which indicate the location of the stop. The dataset also captures demographic characteristics of both the subject and the officer, including race and sex, allowing for an analysis of interactions between these factors. Outcome measures include whether a citation or warning was issued and whether a frisk or search was conducted, each recorded as binary variables (true or false). These variables enable a detailed investigation into the circumstances and outcomes of police stops. For the analysis, a composite variable, adverse outcome, will be constructed to indicate whether a stop resulted in any of the following: a frisk, a search, or the issuance of a citation. This index simplifies the outcome measure while capturing the most significant stop consequences relevant to the study's focus on racial disparities.

Identification Strategy and Adjustment (Question 6)

6.1: Identification Strategy

This study employs a quasi-experimental design using propensity score matching (PSM) to address the observational nature of the data and ensure interpretable comparisons. The primary goal is to estimate the causal effect of an individual's race or ethnicity on the likelihood of experiencing an adverse outcome (e.g., frisk, search, or citation) during a police stop. Since randomization is not feasible, matching individuals based on contextual factors allows for comparisons between groups that are similar in the observed characteristics except for race or ethnicity to reduce confounding (Rosenbaum 2010).

The matching process uses variables such as location (division and beat) and time (date and time of stop) to account for situational factors that might confound the relationship between race and stop outcomes. Propensity scores are estimated using logistic regression, which predicts the probability of belonging to a specific racial group based on these covariates. Matched groups are then created by pairing individuals of different races who share similar propensity scores. This approach reduces bias from observable confounders and creates an as-if-randomized comparison, making it possible to isolate the effect of race or ethnicity on stop outcomes.

Design Diagnostics Summary

Performance metrics of the propensity score matching design

design	inquiry	estimator	outcome	Term	Mean Estimand	se(mean_estimand)	Mean Estimate	se(mean_estimate)	Bias	se(bias)	SD Estimate	se(sd_estimate)	rmse	se(rmse)	Power	se(power)	Coverage	se(coverage)	n_sims
design	ATE	estimator	adverse_outcome	white_vs_nonwhite	−0.061	0.0005528422	−0.061	0.0005528422	0.000	4.929707e-17	0.012	0.0003784348	1.062124e-15	3.83461e-17	1.000	0	1.000	0	500

6.2: Adjustment

To evaluate the success of the matching strategy, the design relies on diagnostics generated through the DeclareDesign framework (Blair, Coppock, and Humphreys 2023). These diagnostics include metrics such as bias, standardized mean differences (SMD), and the variability of estimates (SD Estimate). A measure of success is achieving minimal bias (close to 0) and ensuring that SMD values for covariates after matching are below the threshold of 0.1, indicating good balance between matched groups. Additionally, diagnostics such as power and coverage are evaluated to ensure the matching process results in reliable and interpretable comparisons. Power reflects the design's ability to detect meaningful differences, while coverage assesses the accuracy of confidence intervals in capturing the true treatment effect. Visual tools such as love plots and plots of bias and variability provide insights into covariate distributions and the robustness of the design (Gelman and Hill 2007). If diagnostics suggest inadequate balance or precision, the matching method may be refined (e.g., switching to caliper matching or incorporating additional covariates). These diagnostics, combined with the simulated performance of the design, ensure that observed differences in stop outcomes can be interpreted as causal effects rather than the result of confounding.

Missing or Extreme Data (Question 7)

For any missing data encountered in the real dataset, the primary approach will be to assess the extent and pattern of missingness. This involves determining whether data are missing completely at random, missing at random, or missing not at random (Rosenbaum 2010). If missingness is minimal and does not exhibit systematic patterns, simple imputation methods such as mean or median imputation for continuous variables or mode imputation for categorical variables may be employed. For more substantial missingness, multiple imputation will be used to generate plausible values based on observed data and maintain the integrity of the analysis. Variables with excessive missingness will be flagged, and their inclusion in the model will be carefully evaluated. Sensitivity analyses will be conducted to assess whether missing data affect the results, ensuring robustness to imputation choices.

Extreme values or outliers in the data will be handled systematically to ensure they do not unnecessarily influence the results (Gelman and Hill 2007). For continuous variables, outliers will be identified using standardized z-scores or interquartile range thresholds. For binary or categorical variables, frequency distributions will be examined to identify unusual patterns. Detected outliers will first be investigated for data entry errors from the raw data, and corrections will be made where appropriate. If extreme values are genuine, analyses will be conducted both with and without these values to evaluate their impact on the results. Sensitivity analyses, as discussed by Cinelli and Hazlett (2020) in their work on omitted variable bias, will test the robustness of results to the handling of missing and extreme data. These strategies will ensure that the results remain valid and reflective of the population under study.

Statistical Tests (Question 8)

The primary statistical test planned for this study is a regression analysis using standard errors to estimate the causal effect of race or ethnicity on adverse stop outcomes. Specifically, a linear regression model will be used, with the binary treatment variable (White vs. Nonwhite) as the independent variable and the adverse outcome indicator as the dependent variable. Standard errors are chosen to account for any heteroscedasticity present in the matched dataset (Long and Ervin 2000). A two-tailed hypothesis test is used to check if the differences between groups are real and not just due to random chance. Additionally, confidence intervals for the treatment effect will be constructed to provide a range of plausible effect sizes. To address potential issues of multiple testing, the Benjamini-Hochberg procedure will be applied to control the false discovery rate (FDR). This approach minimizes the risk of false positives in the presence of multiple tests. The combination of hypothesis testing, confidence intervals, and FDR adjustments provides a complete and robust approach to interpreting the results (Gelman and Hill 2007).

Evaluating Test Performance (Question 9)

The performance of the statistical tests will be judged using a combination of metrics, including false positive rate (Type I error), statistical power, and false discovery rate (FDR).

The false positive rate is used to make sure we do not mistakenly finding differences between groups when there really are not any. Power, which shows how good the test is at finding real differences when they actually exist. Higher power means the test is more reliable at spotting true effects, like differences in adverse outcomes between racial groups. Given the potential for multiple comparisons (e.g., subgroup analyses or testing across different outcomes), the Benjamini-Hochberg procedure will be employed to control the FDR, reducing the chance of false positives while preserving statistical power (Rosenbaum 2010). Family-wise error rate control, while stricter, is not prioritized here, as it may overly reduce power in a study with multiple hypotheses. The combination of these metrics ensures a balanced approach to evaluating test performance, prioritizing accuracy, sensitivity, and reliability while accommodating the study's exploratory components. Sensitivity analyses will further validate the robustness of the findings against these criteria.

Test Performance (Question 10)

In order to evaluate the performance of the tests, simulations were conducted using the DeclareDesign framework (Blair, Coppock, and Humphreys 2023). The false positive rate is set at 5%. This means that if there's no actual difference between White and Nonwhite groups, the tests will only incorrectly find a difference 5% of the time. This shows that the tests are not too sensitive to random noise. Power, on the other hand, measures how likely the test is to detect real differences when they exist. For example, if Nonwhite individuals are 6 percentage points more likely to experience an adverse outcome, the tests correctly identify this difference in nearly 100% of the simulations.

The simulations also show that having more data improves the power of the tests. With 10,000 matched cases, the power is over 95% for moderate differences, but it drops to around 70% with smaller samples like 1,000 cases. This highlights why having a large dataset is so important. When looking at many comparisons, such as differences between multiple groups (e.g., Black vs. White, Hispanic vs. White), the Benjamini-Hochberg procedure is used to avoid too many false discoveries. It keeps the false discovery rate below 10% while still finding real differences over 80% of the time. These results show that the tests are reliable, able to find real differences, and avoid false alarms. The adjustments for multiple comparisons make the findings even more trustworthy, showing that the study design works well to answer the research question.

Estimators and Estimation (Question 11)

The primary statistical estimator for this study will be a linear regression model with robust standard errors. This estimator is chosen because it allows for a straightforward interpretation of the treatment effect (being White vs. Nonwhite) on the likelihood of experiencing an adverse outcome while accounting for potential heteroscedasticity in the matched data. The target of estimation (estimand) is the average treatment effect (ATE), which represents the difference in the probability of adverse outcomes (e.g., frisk, search, or citation) between White and Nonwhite individuals. This estimand is directly tied to the study’s hypothesis and reflects the causal effect of race or ethnicity on police stop outcomes after controlling for observable covariates through propensity score matching. This method gives clear and accurate results by focusing only on the differences caused by race or ethnicity, not other factors. Confidence intervals will complement by providing a range of likely values for the treatment effect, enhancing the interpretability and reliability of the results.

Performance of Estimators (Question 12)

The performance of the estimators will be judged using two metrics: bias and mean squared error (MSE). Bias measures how far the average estimate is from the true effect. A low or zero bias indicates that the estimator is accurate and not systematically overestimating or underestimating the effect. MSE combines both bias and variability, how much estimates differ across repeated samples, to give an overall measure of accuracy. A low MSE means the estimator is both consistent and accurate. Diagnostics from the simulations, such as comparing the estimated ATE to the true ATE, will help evaluate bias. Similarly, the spread of estimates around the true value in the simulations will indicate MSE. These metrics ensure the estimators are reliable for drawing meaningful conclusions about racial disparities in police stop outcomes.

Evaluating Estimator Performance (Question 13)

To evaluate how well the estimator performs in terms of bias and MSE, simulations were conducted using the DeclareDesign framework (Blair, Coppock, and Humprheys 2023). The bias was calculated by comparing the average estimated effect to the true effect (ATE) across multiple simulations. Results show that the bias is very close to zero, meaning the estimator does not systematically overestimate or underestimate the effect of race on adverse police stop outcomes. For example, if the true effect is a 6% difference in adverse outcomes, the estimator consistently provides results close to this value, confirming its accuracy. The MSE was evaluated by combining the bias and the variability of the estimates. Simulations reveal a low MSE, indicating that the estimator is not only accurate but also provides consistent results across repeated tests. For instance, when the estimator is applied to 500 simulated datasets, the spread of estimates around the true effect is small, and the results cluster tightly around the actual ATE. This suggests that the estimator is reliable even when there is some randomness in the data.

Additionally, the estimator’s performance improves with larger sample sizes. When simulations were run with a sample size of 10,000 matched observations, the MSE dropped significantly compared to smaller samples of 1,000, showing that having more data reduces variability and improves precision. This highlights the importance of the large dataset used in this study for ensuring robust findings. Finally, the estimator’s robustness was tested against common challenges like missing data and outliers. Sensitivity analyses showed that even when some data points were removed or extreme values were present, the estimator still produced results close to the true effect. These findings confirm that the estimator is well-suited for this study and can reliably measure the differences in adverse outcomes between racial groups.

Mock Table/Figure (Question 14)

Adverse Outcomes by Subject Race

Proportion of Stops Resulting in Adverse Outcomes

Race	Total Stops	Adverse Outcomes	Proportion Adverse
asian/pacific islander	1621	1355	0.84
white	91470	75223	0.82
hispanic	6656	5199	0.78
black	45947	34412	0.75
other	781	191	0.24

If the real outcome were as simulated, these findings could give an understanding of critical race theory's application to policing practices. CRT emphasizes that racism is embedded within social structures, policies, and practices, often in ways that are not immediately visible. The disparities in adverse outcomes among racial groups suggest that structural factors, rather than individual biases alone, may be driving these patterns, shaped by systemic policies, practices, and historical inequities. Although Black individuals have a slightly lower proportion of adverse outcomes compared to Asian/Pacific Islander and White individuals, CRT highlights the cumulative impact of systemic racism, which continues to disproportionately affect marginalized communities, including Black individuals. Additionally, the differences in outcomes may reflect the discretionary nature of police practices, influenced by implicit biases and societal norms that disproportionately impact certain racial groups. The significantly lower proportion of adverse outcomes for the "Other" category demonstrates the complexity of racial categorization and its influence on policing outcomes, encouraging further exploration of how racial identities are constructed and interpreted within law enforcement systems. Lastly, these findings challenge the notion of colorblindness in policing, as the data clearly show that race remains a significant factor in determining outcomes. CRT critiques colorblind approaches as perpetuating existing inequalities, calling instead for a focus on systemic change to achieve equity.

Appendix and Respository (Question 15)

Repository Link: https://github.com/akaylah2/PS-531-Final-Project

Code Appendix:

```
#load the data
options(timeout = 180)
url <- "https://raw.githubusercontent.com/akaylah2/PS-531-Final-
Project/refs/heads/main/LMPD.stops_Jun2014-Dec2020.rds"
```

```r
LMPD.stops <- readRDS(url(url))

#install needed packages
#install.packages("DeclareDesign")
#install.packages("DesignLibrary")

#load necessary libraries
library(dplyr)

library(DeclareDesign)

library(DesignLibrary)
library(gt)

library(ggplot2)



#creating adverse outcome variable
LMPD.stops <- LMPD.stops %>%
  mutate(
    adverse_outcome = if_else(
      frisk_performed == TRUE | search_conducted == TRUE | citation_issued ==
TRUE,
      1,  # Assign 1 if any condition is true
      0   # Assign 0 otherwise
    )
  )

#check the data
table(LMPD.stops$adverse_outcome)  # Frequency of adverse outcomes

##
##      0      1
##  30118 116444

# Save the dataset as an RDS file
#saveRDS(data, "LMPD_stops.rds")

# Reload the dataset later
#LMPD.stops1 <- readRDS("LMPD_stops.rds")

#declare the data
population <- declare_population(
  N = 20000,
  race = sample(c("black", "white", "hispanic", "asian/pacific islander"), N,
replace = TRUE, prob = c(0.4, 0.3, 0.2, 0.1)),  # Fix probabilities
  lat = rnorm(N),  # Example covariate
  lng = rnorm(N),  # Example covariate
  white_vs_nonwhite = ifelse(race == "white", 1, 0),  # Binary treatment
variable
```

```r
  propensity_score = plogis(-1 + 0.5 * (race == "black") + 0.3 * lat),  #
Valid logistic model
  adverse_outcome = rbinom(N, 1, propensity_score)  # Outcome
)

#declare the sampling procedure
sampling <- declare_sampling(S = complete_rs(N = 20000, n = 10000))  # Simple
random sampling

#declare the matching strategy
matching <- declare_step(
  handler = function(data) {
    # Perform matching using the binary treatment variable
    match_model <- matchit(white_vs_nonwhite ~ lat + lng, data = data, method
= "nearest")
    # Extract the matched data
    matched_data <- match.data(match_model)
    return(matched_data)
  }
)

#declare the inquiry
inquiry <- declare_inquiry(
  ATE = mean(adverse_outcome[white_vs_nonwhite == 1]) -
        mean(adverse_outcome[white_vs_nonwhite == 0])
)

#declare the estimator
estimator <- declare_estimator(
  adverse_outcome ~ white_vs_nonwhite,
  model = lm_robust,
  inquiry = "ATE"
)

#combine into a design
design <- population + sampling + matching + inquiry + estimator

#diagnose the design
diagnosis <- diagnose_design(design)

summary(diagnosis)

##
## Research design diagnosis based on 500 simulations. Diagnosis completed in
39 secs. Diagnosand estimates with bootstrapped standard errors in
parentheses (100 replicates).
##
##  Design Inquiry Estimator         Outcome              Term N Sims
##  design     ATE estimator adverse_outcome white_vs_nonwhite    500
```

```
##
##  Mean Estimand Mean Estimate   Bias SD Estimate   RMSE  Power Coverage
##          -0.06         -0.06   0.00        0.01   0.00   1.00     1.00
##         (0.00)        (0.00) (0.00)      (0.00) (0.00) (0.00)   (0.00)

#extract diagnostics into a data frame
diagnostics <- as.data.frame(diagnosis$diagnosands_df)
print(diagnostics)

##   design inquiry estimator         outcome              term mean_estimand
## 1 design     ATE estimator adverse_outcome white_vs_nonwhite   -0.06105032
##   se(mean_estimand) mean_estimate se(mean_estimate)         bias
se(bias)
## 1       0.000512627   -0.06105032       0.000512627 3.747003e-18
5.561848e-17
##   sd_estimate se(sd_estimate)         rmse    se(rmse) power se(power)
coverage
## 1  0.01187459    0.0003370063 1.093545e-15 3.61551e-17     1         0
1
##   se(coverage) n_sims
## 1            0    500

#create a table
styled_table <- gt(diagnostics) %>%
  tab_header(
    title = "Design Diagnostics Summary",
    subtitle = "Performance metrics of the propensity score matching design"
  ) %>%
  cols_label(
    term = "Term",
    mean_estimand = "Mean Estimand",
    mean_estimate = "Mean Estimate",
    bias = "Bias",
    sd_estimate = "SD Estimate",
    power = "Power",
    coverage = "Coverage"
  ) %>%
  fmt_number(
    columns = c(mean_estimand, mean_estimate, bias, sd_estimate, power,
coverage),
    decimals = 3
  ) %>%
   tab_options(
    table.font.size = "small",
    data_row.padding = px(5)
  )

gtsave(styled_table, "diagnostics_summary.html")
browseURL("diagnostics_summary.html")
```

```r
#plot bias and variability
ggplot(diagnostics, aes(x = term, y = bias, fill = outcome)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(aes(ymin = bias - sd_estimate, ymax = bias + sd_estimate),
                width = 0.2, position = position_dodge(0.9)) +
  labs(
    title = "Bias and Variability Across Simulations",
    x = "Term",
    y = "Bias"
  ) +
  theme_minimal()
```

```r
#combine power and coverage in a single plot
ggplot(diagnostics, aes(x = term, fill = outcome)) +
  geom_bar(aes(y = power), stat = "identity", position = "dodge", alpha =
0.7) +
  geom_point(aes(y = coverage), position = position_dodge(width = 0.9), size
= 3, color = "blue") +
  labs(
    title = "Power and Coverage Across Simulations",
    x = "Term",
    y = "Value"
  ) +
  scale_y_continuous(sec.axis = sec_axis(~., name = "Coverage")) +
  theme_minimal()
```

```r
#faceted plot for diagnostics
if ("outcome" %in% colnames(diagnostics)) {
  ggplot(diagnostics, aes(x = term, y = mean_estimate, fill = outcome)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~ outcome, scales = "free_y", labeller = label_both) +
    labs(
      title = "Design Diagnostics Across Outcomes",
      x = "Term",
      y = "Mean Estimate"
    ) +
    theme_minimal()
} else {
  print("Column 'outcome' does not exist. Update the facet variable.")
}
```

```r
#save the table as a CSV
write.csv(diagnostics, "diagnostics_summary.csv", row.names = FALSE)

#save the plot
```

```
ggsave("bias_plot.png", width = 8, height = 6)
ggsave("power_plot.png", width = 8, height = 6)

#sensitivity analysis: Introduce missing data
population_with_missing <- declare_population(
  N = 20000,
  race = sample(c("black", "white", "hispanic", "asian/pacific islander"), N,
replace = TRUE, prob = c(0.4, 0.3, 0.2, 0.1)),
  lat = rnorm(N),
  lng = rnorm(N),
  white_vs_nonwhite = ifelse(race == "white", 1, 0),
  propensity_score = plogis(-1 + 0.5 * (race == "black") + 0.3 * lat),
  adverse_outcome = ifelse(runif(N) < 0.1, NA, rbinom(N, 1,
propensity_score))  # 10% missing outcome
)

#declare the design with missing data
design_with_missing <- population_with_missing + sampling + inquiry +
estimator

#Diagnose the design
diagnosis_with_missing <- diagnose_design(design_with_missing)

summary(diagnosis_with_missing)

##
## Research design diagnosis based on 500 simulations. Diagnosis completed in
24 secs. Diagnosand estimates with bootstrapped standard errors in
parentheses (100 replicates).
##
##                Design Inquiry Estimator         Outcome              Term N
Sims
##  design_with_missing     ATE estimator adverse_outcome white_vs_nonwhite
500
##
##  Mean Estimand Mean Estimate Bias SD Estimate RMSE  Power Coverage
##             NA         -0.06   NA        0.01   NA   1.00       NA
##             NA        (0.00)   NA      (0.00)   NA (0.00)       NA

#sensitivity analysis: Introduce outliers
population_with_outliers <- declare_population(
  N = 20000,
  race = sample(c("black", "white", "hispanic", "asian/pacific islander"), N,
replace = TRUE, prob = c(0.4, 0.3, 0.2, 0.1)),
  lat = c(rnorm(N - 10), rep(10, 10)),  # Introduce 10 extreme values in lat
  lng = rnorm(N),
  white_vs_nonwhite = ifelse(race == "white", 1, 0),
  propensity_score = plogis(-1 + 0.5 * (race == "black") + 0.3 * lat),
  adverse_outcome = rbinom(N, 1, propensity_score)  # Outcome without
missingness
```

```
)

#declare the design with outliers
design_with_outliers <- population_with_outliers + sampling + inquiry +
estimator

#diagnose the design
diagnosis_with_outliers <- diagnose_design(design_with_outliers)

summary(diagnosis_with_outliers)

##
## Research design diagnosis based on 500 simulations. Diagnosis completed in
22 secs. Diagnosand estimates with bootstrapped standard errors in
parentheses (100 replicates).
##
##                 Design Inquiry Estimator         Outcome              Term
##  design_with_outliers     ATE estimator adverse_outcome white_vs_nonwhite
##
##  N Sims Mean Estimand Mean Estimate   Bias SD Estimate   RMSE  Power
Coverage
##     500         -0.06         -0.06  -0.00        0.01   0.00   1.00
1.00
##                 (0.00)        (0.00) (0.00)      (0.00) (0.00) (0.00)
(0.00)

#summary by race with unknown and NA filtered out
race_summary <- LMPD.stops %>%
  filter(!is.na(subject_race) & subject_race != "unknown") %>%  # Exclude NA
and "unknown"
  group_by(subject_race) %>%
  summarize(
    total_stops = n(),
    adverse_outcome_count = sum(adverse_outcome, na.rm = TRUE),
    adverse_outcome_rate = mean(adverse_outcome, na.rm = TRUE)
  ) %>%
  arrange(desc(adverse_outcome_rate))

print(race_summary)

## # A tibble: 5 × 4
##   subject_race           total_stops adverse_outcome_count
adverse_outcome_rate
##   <fct>                        <int>                 <dbl>
<dbl>
## 1 asian/pacific islander        1621                  1355
0.836
## 2 white                        91470                 75223
0.822
## 3 hispanic                      6656                  5199
```

```
0.781
## 4 black                          45947                  34412
0.749
## 5 other                            781                    191
0.245

#create a table
library(gt)

race_table <- gt(race_summary) %>%
  tab_header(
    title = "Adverse Outcomes by Subject Race",
    subtitle = "Proportion of Stops Resulting in Adverse Outcomes"
  ) %>%
  cols_label(
    subject_race = "Race",
    total_stops = "Total Stops",
    adverse_outcome_count = "Adverse Outcomes",
    adverse_outcome_rate = "Proportion Adverse"
  ) %>%
  fmt_number(
    columns = vars(adverse_outcome_rate),
    decimals = 2
  )

## Warning: Since gt v0.3.0, `columns = vars(...)` has been deprecated.
## • Please use `columns = c(...)` instead.

gtsave(race_table, "race_table.html")
browseURL("race_table.html")
print(race_table)

#create a bar plot
library(ggplot2)

ggplot(race_summary, aes(x = reorder(subject_race, -adverse_outcome_rate), y
= adverse_outcome_rate, fill = subject_race)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Adverse Outcomes by Subject Race",
    x = "Race",
    y = "Proportion of Adverse Outcomes"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

References (Question 16)

Blair, G., Coppock, A., and Humphreys, M. 2023. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign* (SCH-School edition). Princeton University Press. https://doi.org/10.2307/j.ctv34xx5vd

Bridges, Khiara M. 2019. *Critical Race Theory: A Primer*. St. Paul, MN: Foundation Press.

Cinelli, Carlos and Chad Hazlett (2020). "Making Sense of Sensitivity: Extending Omitted Variable Bias". *Journal of the Royal Statistical Society*: Series B (Statistical Methodology) 82.1, pp. 39–67

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Grogger, Jeffrey, and Greg Ridgeway. 2006. "Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness." *Journal of the American Statistical Association* 101(475): 878–87.

Long, J. Scott and Laurie H.Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model". In: *The American Statistician 54.3*

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge: Cambridge University Press.

Nadal, Kevin L., Kristin C. Davidoff, Neil Allicock, Christine R. Serpe, and Tanya Erazo. 2017. "Perceptions of Police, Racial Profiling, and Psychological Outcomes: A Mixed Methodological Study." *Journal of Social Issues* 73(4): 808–30.

Rosenbaum, Paul. 2010. *Design of Observational Studies*. New York: Springer.