Project Proposal

Due November 17 at 11:59pm

Jenny Wu, Rishika Randev, Shiyue Zhou, Uzoma Uwazurike

Load Packages

Dataset 1 (top choice)

Data source: County Health Rankings & Roadmaps https://www.countyhealthrankings.org/health-data/north-carolina/data-and-resources)

Secondary dataset: https://data.census.gov/table/ACSDT1Y2018.B25103?q=property

Brief description: The source of the datasets includes health factor and health outcome data for each county in the US by state in 2024. These datasets are provided as part of a program by the University of Wisconsin Population Health Institute and combines data from other sources, such as the Behavioral Risk Factor Surveillance System & National Center for Health Statistics. We are compiling the datasets for seven states in seven different regions of the US to create one dataset (North Carolina, Texas, Alabama, Colorado, Minnesota, Washington, & New York). We will also be merging in a second American Community Survey dataset to obtain property tax data for our second research question.

Research question 1: What is the relationship between residential segregation (continuous variable), percentage of the county population with limited access to healthy foods (continuous variable), and percentage of adults who are uninsured (continuous variable), and average life expectancy (continuous variable)? How is the relationship between residential segregation and life expectancy affected by the region of the US that the county / state is in (categorical variable that we will create called Region)? The outcome variable is average life expectancy. The predictors are residential segregation, percentage with limited access to healthy foods, and percentage of adults who are uninsured.

Average life expectancy = average number of years a person is expected to live (data from National Center for Health Statistics - Natality and Mortality Files; Census Population Estimates Program, 2019-2021).

Residential segregation = Index of dissimilarity where higher values indicate greater residential segregation between Black and white county residents (data from American Community Survey 5-year estimate, 2018-2022).

Percentage limited access to healthy foods = Percentage of population who are low-income and do not live close to a grocery store (data from USDA Food Environment Atlas, 2019).

Percentage adults uninsured = Percentage of adults under age 65 without health insurance (data from Small Area Health Insurance Estimates, 2021).

Research question 2: What is the relationship between residential segregation (continuous variable), property tax (continuous variable), and median household income (continuous variable), on school funding adequacy (categorical variable that we will create)?

The outcome variable is school funding adequacy, with 0=inadequate funding and 1=adequate funding. The predictors are residential segregation, property tax, and median household income.

School funding adequacy = The average gap in dollars between actual and required spending per pupil among public school districts. Required spending is an estimate of dollars needed to achieve U.S. average test scores in each district. (data from School Finance Indicators Database, 2021).

Median property tax = Median property tax value for each county between 2018-2022 (data from IPUMS).

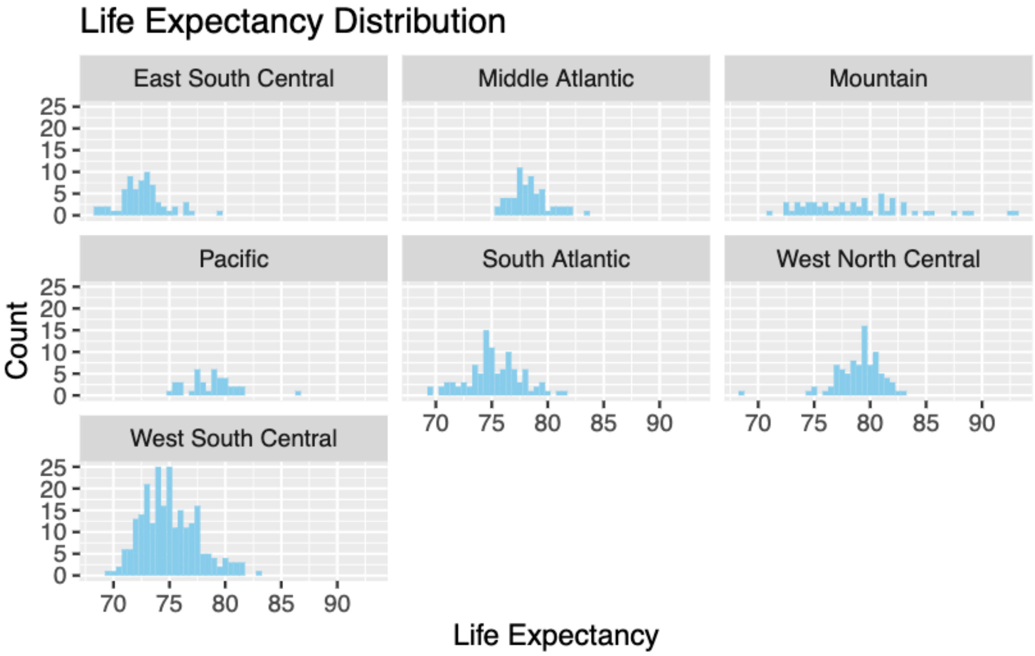
Median household income = The income where half of households in a county earn more and half of households earn less (data from Small Area Income and Poverty Estimates; American Community Survey, 5-year estimates, 2022 & 2018-2022).

Load the datasets and prepare final dataset (code is hidden to reduce proposal length):

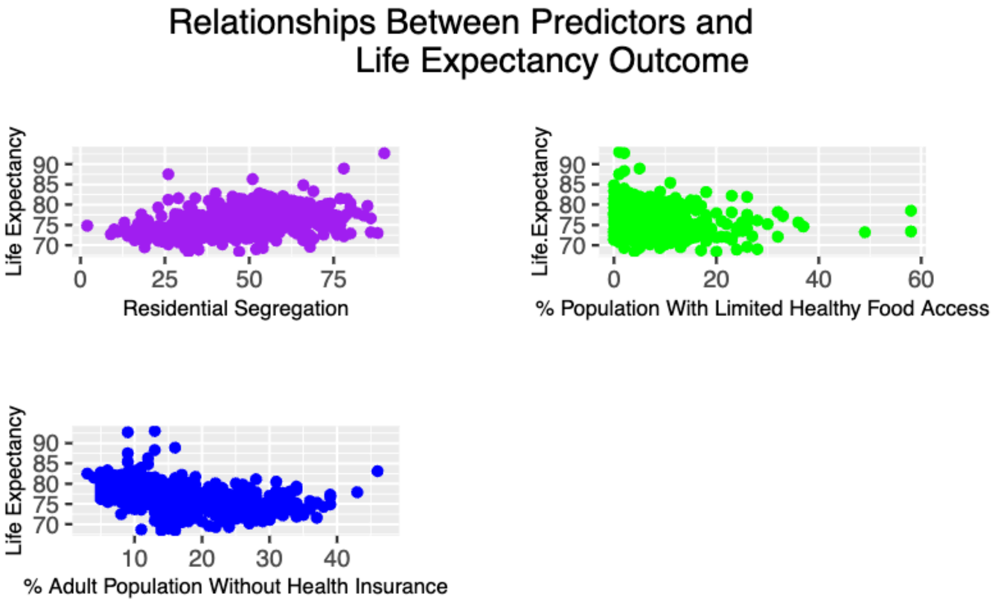
Provide a ‘glimpse():

```
Rows: 673
Columns: 13
$ FIPS                               <int> 1001, 1003, 1005, 1007, 1009, 101~
$ State                              <chr> "Alabama", "Alabama", "Alabama", ~
$ County                             <chr> "Autauga", "Baldwin", "Barbour", ~
$ Life.Expectancy                    <dbl> 75.3, 76.7, 72.4, 72.3, 73.4, 72.~
$ Residential.Segregation.Index      <int> 29, 40, 19, 32, 63, 35, 24, 46, 3~
$ Percent.Limited.Access.Healthy.Foods <int> 13, 8, 10, 0, 3, 32, 7, 15, 8, 5,~
$ Percent.Uninsured.Adults           <int> 12, 13, 16, 14, 16, 16, 15, 17, 1~
$ Spending.per.Pupil                 <int> 9098, 11638, 8807, 10694, 9138, 1~
$ School.Funding.Adequacy            <int> -3607, -537, -23627, -6971, -2789~
$ Median.Household.Income            <int> 70148, 71704, 41151, 54309, 60553~
$ Region                         <fct> East South Central, East South Ce~
$ School.Funding.Cat             <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ Median.Prop.Tax                <dbl> 531, 796, 398, 286, 464, 394, 315~
```

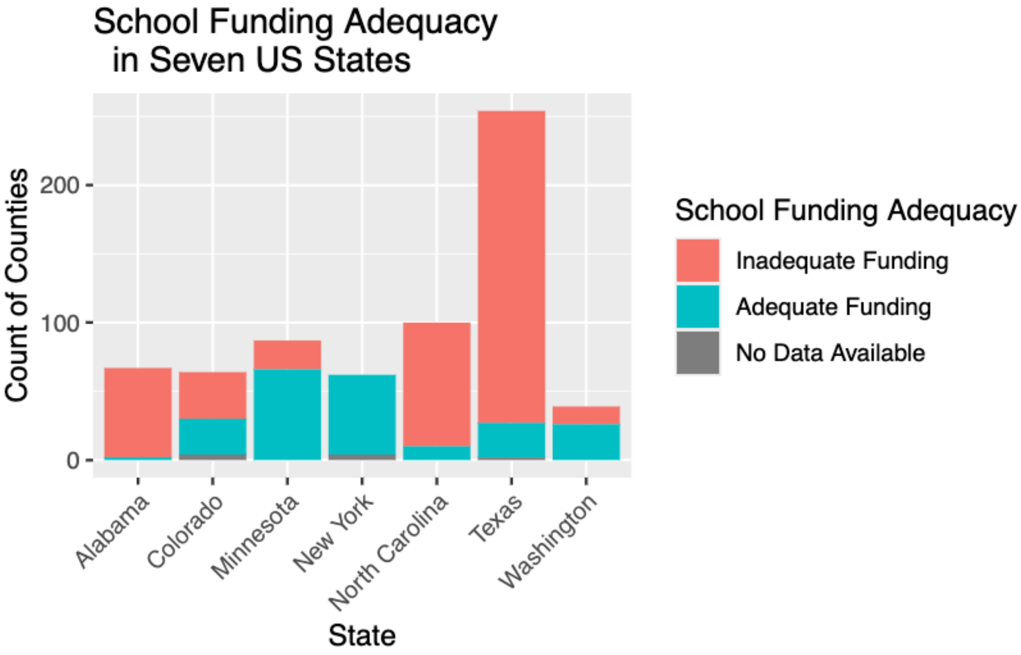
Exploratory Plots: Q1 Outcome Variable (Life Expectancy):



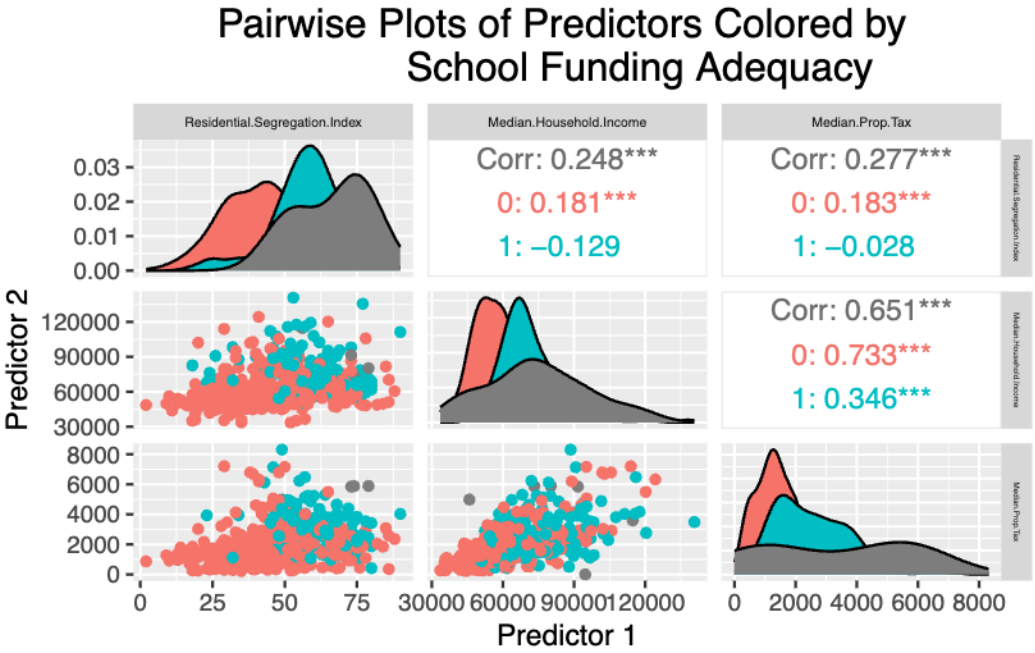
Q1 Relationship of Interest:



Q2 Outcome Variable (School Funding Adequacy):



Q2 Relationship of Interest:



The above figure is a pairwise scatterplot for each pair of predictor variables (residential segregation, property tax, and household income), with the points being colored by the school funding adequacy outcome. Along the diagonal are histograms showing the distribution of each individual predictor, separated by each level of the school funding adequacy outcome variable (red = 0, or inadequate funding; blue = 1, or adequate funding; gray = NA values). The upper half shows the correlation coefficient between each pair of predictors, and also the specific correlation coefficients when the data is separated out by each level of the outcome.

Dataset 2

Data source: NFL Kaggle 2025 Play Dataset https://www.kaggle.com/competitions/nfl-big-data-bowl-2025/data?select=player_play.csv

Brief description: The Play Data Dataset provides detailed play-by-play information about football games, capturing various aspects of gameplay, team dynamics, and player actions. This dataset includes variables that tell us more about what influences game play within a football game.

Research question 1: Does the distance that a quarterback drops back after the snap and the dropback type affect the probability of a pass completion?

Outcome Variable: `passResult` (Binary: Complete/Incomplete), indicating whether the pass was successfully completed.

Primary Predictors:

- `dropbackDistance` (Continuous).The distance the QB dropped back (yards) behind the center after the snap (numeric)
- `dropbackType` (Categorical).The type of drop back after the snap by the QB (Traditional, Designed Rollout, Scramble, Scramble Rollout, Designed Rollout Left, Designed Rollout Right, Scramble Rollout Left, Scramble Rollout Right, Designed Run, QB Draw, Rollout, text) (categorical)

Potential Interaction Terms:

- `dropbackDistance* dropbackType` : Different types of dropbacks (e.g., "Rollout" vs. "Traditional") may influence the effectiveness of a specific dropback distance.

Research question 2: How does the game situation (down, quarter, and opponent score) affect home team's win probability change after a play?

Outcome Variable: `homeTeamWinProbabilityAdded` (Continuous), indicating how much the home team's win probability changes after a given play.

Primary Predictors:

- `quarter` (Discrete), indicating whether the play is in the first, second, third, or fourth quarter, or during overtime.
- `down` (Discrete), indicating whether the play is on a first, second, third, or fourth down.
- `preSnapVisitorScore` (Discrete), indicating the score of the oponent team prior to the current play.

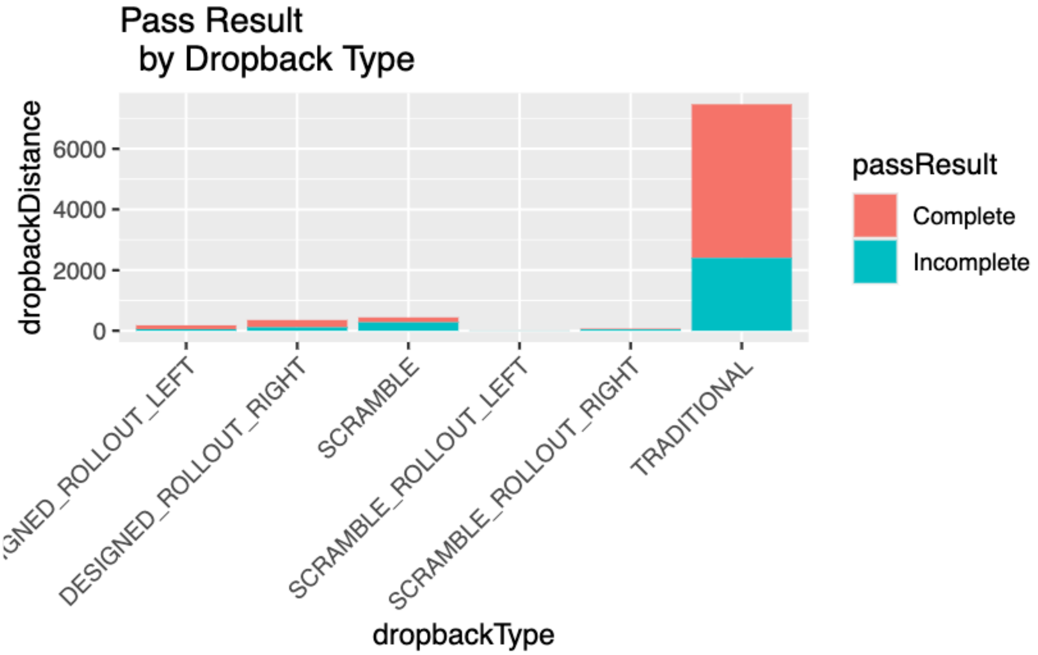
Load the data and provide a `glimpse()`:

```
Rows: 16,124
Columns: 50
$ gameId                           <int> 2022102302, 2022091809, 2022103004, 2~
$ playId                           <int> 2655, 3698, 3146, 348, 2799, 2314, 38~
$ playDescription                  <chr> "(1:54) (Shotgun) J.Burrow pass short~
$ quarter                          <int> 3, 4, 4, 1, 3, 3, 4, 4, 4, 2, 1, 4, 2~
$ down                             <int> 1, 1, 3, 2, 2, 2, 1, 3, 3, 3, 3, 2, 1~
$ yardsToGo                        <int> 10, 10, 12, 10, 8, 6, 10, 12, 12, 8, ~
$ possessionTeam                   <chr> "CIN", "CIN", "HOU", "KC", "BAL", "DE~
$ defensiveTeam                    <chr> "ATL", "DAL", "TEN", "TEN", "TB", "SE~
$ yardlineSide                     <chr> "CIN", "CIN", "HOU", "TEN", "TB", "DE~
$ yardlineNumber                   <int> 21, 8, 20, 23, 27, 29, 40, 28, 35, 35~
$ gameClock                        <chr> "01:54", "02:13", "02:00", "09:28", "~
$ preSnapHomeScore                 <int> 35, 17, 3, 0, 10, 15, 26, 16, 28, 6, ~
$ preSnapVisitorScore              <int> 17, 17, 17, 0, 10, 31, 3, 26, 38, 7, ~
$ playNullifiedByPenalty           <chr> "N", "N", "N", "N", "N", "N", "N", "N~
$ absoluteYardlineNumber           <int> 31, 18, 30, 33, 37, 39, 50, 82, 45, 4~
$ preSnapHomeTeamWinProbability    <dbl> 0.982017488, 0.424356237, 0.006291237~
$ preSnapVisitorTeamWinProbability <dbl> 0.017982512, 0.575643763, 0.993708763~
$ expectedPoints                   <dbl> 0.7193135, 0.6077456, -0.2914852, 4.2~
$ offenseFormation                 <chr> "EMPTY", "EMPTY", "SHOTGUN", "SHOTGUN~
$ receiverAlignment                <chr> "3x2", "3x2", "2x2", "2x2", "3x1", "3~
$ playClockAtSnap                  <int> 10, 9, 12, 11, 8, 15, 18, 2, 3, 12, 1~
$ passResult                       <chr> "C", "C", "C", "C", "", "", "", "", "~
$ passLength                       <int> 6, 4, -4, -6, NA, NA, NA, NA, -6, 15,~
$ targetX                          <dbl> 36.69, 20.83, 26.02, 38.95, NA, NA, N~
$ targetY                          <dbl> 16.51, 20.49, 17.56, 14.19, NA, NA, N~
$ playAction                       <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FAL~
$ dropbackType                     <chr> "TRADITIONAL", "TRADITIONAL", "TRADIT~
$ dropbackDistance                 <dbl> 2.40, 1.14, 3.20, 3.02, 2.03, NA, NA,~
$ passLocationType                 <chr> "INSIDE_BOX", "INSIDE_BOX", "INSIDE_B~
$ timeToThrow                      <dbl> 2.990, 1.836, 2.236, 2.202, NA, NA, N~
$ timeInTackleBox                  <dbl> 2.990, 1.836, 2.236, 2.202, NA, NA, N~
$ timeToSack                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ passTippedAtLine                 <lgl> FALSE, FALSE, FALSE, FALSE, NA, NA, N~
$ unblockedPressure                <lgl> FALSE, FALSE, FALSE, FALSE, NA, NA, N~
$ qbSpike                          <lgl> FALSE, FALSE, FALSE, FALSE, NA, NA, N~
$ qbKneel                          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
$ qbSneak                          <lgl> NA, NA, NA, NA, FALSE, FALSE, FALSE, ~
$ rushLocationType                 <chr> NA, NA, NA, NA, "INSIDE_LEFT", "INSID~
$ penaltyYards                     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ prePenaltyYardsGained            <int> 9, 4, 6, 4, -1, 3, 5, -1, 0, 15, 0, 0~
$ yardsGained                      <int> 9, 4, 6, 4, -1, 3, 5, -1, 0, 15, 0, 0~
$ homeTeamWinProbabilityAdded      <dbl> 4.633843e-03, 2.846926e-03, 2.047173e~
$ visitorTeamWinProbilityAdded     <dbl> -4.633843e-03, -2.846926e-03, -2.0471~
$ expectedPointsAdded              <dbl> 0.70271669, -0.24050862, -0.21848040,~
$ isDropback                       <lgl> TRUE, TRUE, TRUE, TRUE, FALSE, FALSE,~
$ pff_runConceptPrimary            <chr> NA, NA, NA, NA, "MAN", "MAN", "INSIDE~
$ pff_runConceptSecondary          <chr> NA, NA, NA, NA, "READ OPTION", NA, NA~
$ pff_runPassOption                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ pff_passCoverage                 <chr> "Cover-3", "Quarters", "Quarters", "Q~
$ pff_manZone                      <chr> "Zone", "Zone", "Zone", "Zone", "Man"~
```

Exploratory Plots: RQ1: Does the distance that a quarterback drops back after the snap affect the probability of a pass completion?

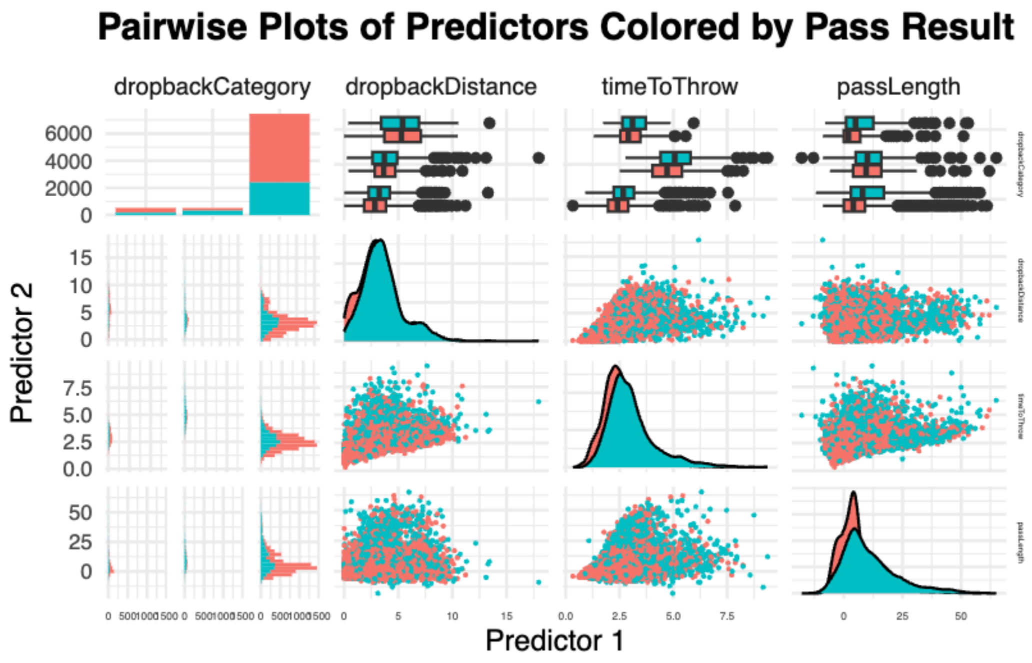
Q1 Outcome Plot:

```
Rows: 8,512
Columns: 5
$ dropbackDistance <dbl> 2.40, 1.14, 3.20, 3.02, 1.78, 5.16, 2.27, 4.76, 3.21,~
$ dropbackType     <fct> TRADITIONAL, TRADITIONAL, TRADITIONAL, TRADITIONAL, T~
$ timeToThrow      <dbl> 2.990, 1.836, 2.236, 2.202, 1.568, 3.203, 2.130, 2.97~
$ passLength       <int> 6, 4, -4, -6, -6, 15, 5, 12, 2, 11, -3, 17, 15, 28, 4~
$ passResult       <fct> C, C, C, C, I, C, I, C, C, I, C, C, C, C, I, C, C, C,~
```



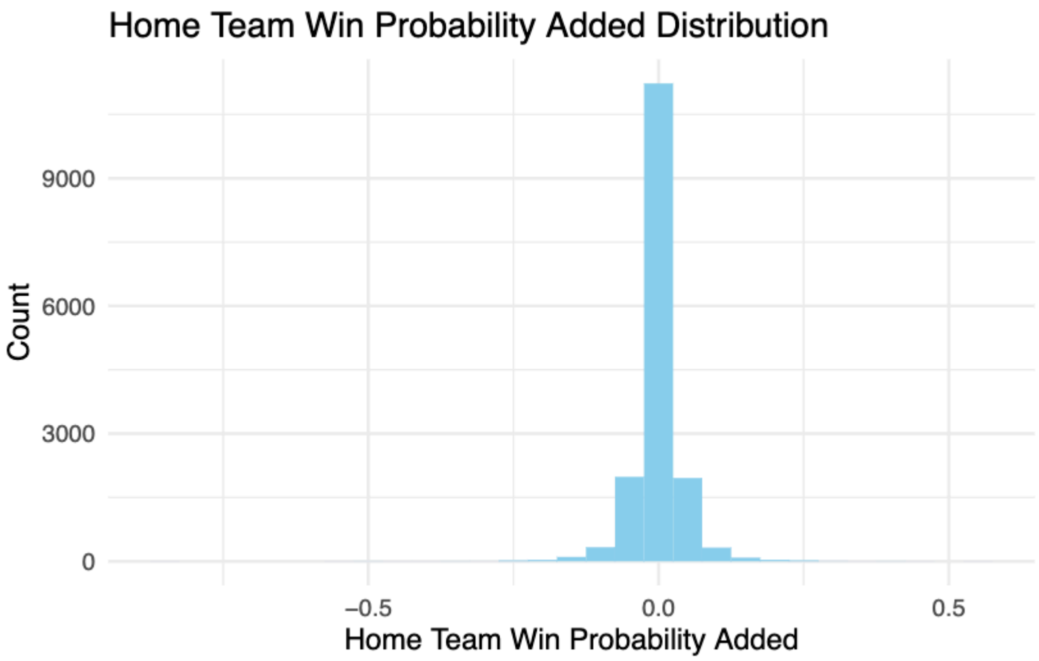
Q1 Relationship of Interest:

```
 DESIGNED_ROLLOUT_LEFT DESIGNED_ROLLOUT_RIGHT               SCRAMBLE 
                   178                    351                    438 
 SCRAMBLE_ROLLOUT_LEFT SCRAMBLE_ROLLOUT_RIGHT            TRADITIONAL 
                    14                     67                   7464 
```

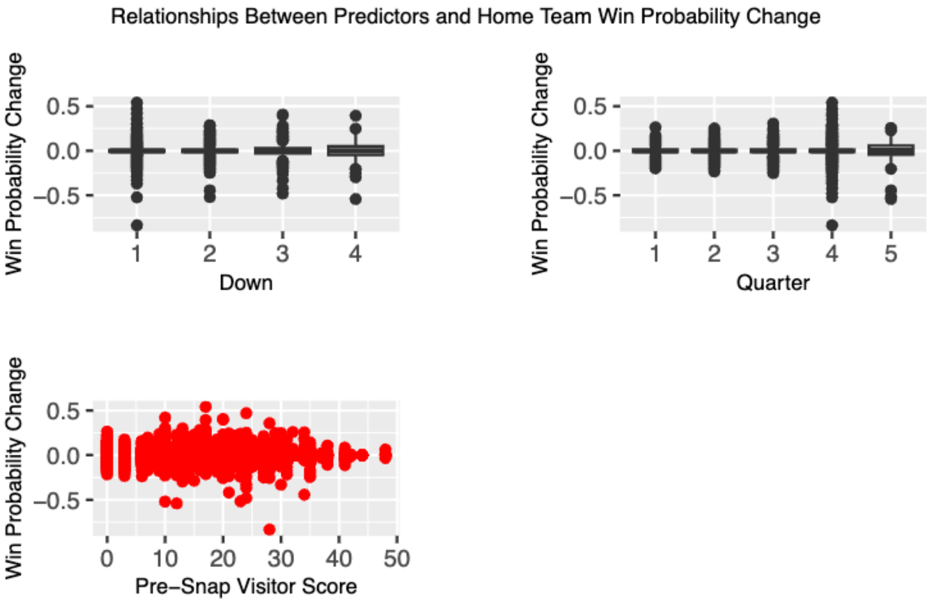


RQ2: How does the game situation affect win probability change?

Q2 Outcome Plot:



Q2 Relationship of Interest:



Team Charter

When will you meet as a team to work on the project components? Will these meetings be held in person or virtually?

We will meet virtually on Tuesday between 8 and 9 pm.

What is your group policy on missing team meetings (e.g., how much advance notice should be provided)?

If a team member will be missing a meeting, they should provide around 10-12 hours of advance notice (the morning of).

How will your team communicate (email, Slack, text messages)? What is your policy on appropriate response time (within a certain number of hours? Nights/weekends?)?

We will be using WhatsApp primarily for communication, with Slack as our secondary chat. We will send Zoom invites via email to reserve our weekly meeting time (Tuesdays 8-9 pm). Ideally, team members should respond to or acknowledge messages within 1-3 hours on weekdays and 12 hours on weekends/during the night.