Project Proposal

Due November 17 at 11:59pm

Nathan Bush, Lilah DuBoff, Syed Huma Shah, and Atreya Tadepalli

Load Packages

```
options(repos = c(CRAN = "https://cran.rstudio.com"))
library(tidyverse)
library(dplyr)
library(ggplot2)
install.packages("corrplot")
```

```
The downloaded binary packages are in
    /var/folders/gk/rvx2k2rs1yvbbbfgbclm80600000gn/T//Rtmpcou1gm/downloaded_packages
```

```
library(corrplot)


rm(list = ls())
```

Dataset 1

Data source: Mendeley Data, https://data.mendeley.com/datasets/wj9rwkp9c2/1

Brief description: The data were collected from an Iraqi laboratory within the Medical City Hospital, the Specialized Center for Endocrinology, and the Diabetes-Al-Kindy Teaching Hospital. The dataset is composed of patient files, inlcuding medical information, laboratory analysis, and diagnoses.

Long description: This dataset contains information about individuals who are classified as not having diabetes, are prediabetic, or have been diagnosed with diabetes. Variables include:

1. Patient ID
2. Blood Sugar Level
3. Age
4. Gender
5. Creatinine ratio(Cr)
6. Body Mass Index (BMI)
7. Urea
8. Cholesterol (Chol)
9. Fasting lipid profile including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol
10. HBA1C
11. Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Pre-Diabetic).

The dataset allows for an investigation into the factors that contribute to the risk of developing diabetes, and provides an opportunity to uncover patterns and relationships that could inform both clinical practice and public health interventions. The variables can be classified into three types: biochemical markers (blood sugar, urea, creatinine), lipid profiles, and HbA1c levels. Together, these variables are essential for understanding the metabolic profile of individuals across different diabetes classifications. The age and gender categories can help to identify the most-at risk groups.

Research question 1: To what extent do demographic factors, such as age and gender, influence a patient's likelihood of being classified as diabetic, pre-diabetic, or non-diabetic? Additionally, how do these demographic characteristics interact with other clinical variables, such as BMI, blood sugar levels, or lipid profiles, in determining diabetes risk and progression?

Outcome Variable: CLASS (Diabetes status of patient)

Type: Ordinal Categorical Variable

Predictor Variables: Age, Gender, potentially other interaction variables

Inference Goal: This question seeks to determine if either age or gender are statistically significant variables when predicting the diabetes class of a patient, and if age interacting with other clinical variables has an impact on classification.

Research question 2: Do health information statistics like blood sugar level, cholesterol, or lipids profiles have an impact on whether a patient is classified as diabetic, pre-diabetic, or non-diabetic?

Outcome Variable Name: *CLASS* (Diabetes status of patient)

Type: Ordinal Categorical Variable

Predictor Variables:

Urea - Urea amount (Continuous, mmol/L)

Chol - Cholesterol measurement (Continuous, mg/dL)

BMI - Body Mass Index (Continuous)

HBA1C - Hemoglobin A1C; long-term blood sugar levels (Continuous, %)

Cr - Creatinine ratio (Continuous, mg/g)

TG - Triglycerides (Continuous, mg/dL)

HDL - High-density Lipoprotein Cholesterol (Continuous, mg/dL)

VLDL - Very Low-density Lipoprotein Cholesterol (Continuous, mg/dL)

LDL - Low-density Lipoprotein Cholesterol (Continuous, mg/dL)

Description: This question seeks to determine if the diabetes class of a patient can be accurately predicted using the health indicators available

Research Question 1 EDA and Plots

```r
diabetes_df <- read.csv("https://raw.githubusercontent.com/lilah-duboff/Stats_Final_Project/

glimpse(diabetes_df)
```

```
Rows: 1,000
Columns: 14
$ ID        <int> 502, 735, 420, 680, 504, 634, 721, 421, 670, 759, 636, 788, ~
$ No_Pation <int> 17975, 34221, 47975, 87656, 34223, 34224, 34225, 34227, 3422~
$ Gender    <chr> "F", "M", "F", "F", "M", "F", "F", "M", "M", "F", "F", "F", ~
$ AGE       <int> 50, 26, 50, 50, 33, 45, 50, 48, 43, 32, 31, 33, 30, 45, 50, ~
$ Urea      <dbl> 4.7, 4.5, 4.7, 4.7, 7.1, 2.3, 2.0, 4.7, 2.6, 3.6, 4.4, 3.3, ~
$ Cr        <int> 46, 62, 46, 46, 46, 24, 50, 47, 67, 28, 55, 53, 42, 54, 39, ~
$ HbA1c     <dbl> 4.9, 4.9, 4.9, 4.9, 4.9, 4.0, 4.0, 4.0, 4.0, 4.0, 4.2, 4.0, ~
$ Chol      <dbl> 4.2, 3.7, 4.2, 4.2, 4.9, 2.9, 3.6, 2.9, 3.8, 3.8, 3.6, 4.0, ~
$ TG        <dbl> 0.9, 1.4, 0.9, 0.9, 1.0, 1.0, 1.3, 0.8, 0.9, 2.0, 0.7, 1.1, ~
$ HDL       <dbl> 2.4, 1.1, 2.4, 2.4, 0.8, 1.0, 0.9, 0.9, 2.4, 2.4, 1.7, 0.9, ~
$ LDL       <dbl> 1.4, 2.1, 1.4, 1.4, 2.0, 1.5, 2.1, 1.6, 3.7, 3.8, 1.6, 2.7, ~
$ VLDL      <dbl> 0.5, 0.6, 0.5, 0.5, 0.4, 0.4, 0.6, 0.4, 1.0, 1.0, 0.3, 1.0, ~
$ BMI       <dbl> 24, 23, 24, 24, 21, 21, 24, 24, 21, 24, 23, 21, 22, 23, 24, ~
$ CLASS     <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", ~
```

```r
diabetes_df |> count(CLASS)
```

```
  CLASS   n
1     N 102
2    N    1
3     P  53
4     Y 840
5    Y    4
```

```r
diabetes_df |> count(Gender)
```

```
  Gender   n
1      F 434
2      M 565
3      f   1
```

As we can see from these counts tables, there are a few issues with how the data was input into the dataset. Notably, there are duplicate labels for how the diabetes classification was input - this is likely due to white-space, so in the next chunk, we've removed it, and checked the table counts again. In the gender table, there is one entry where a female patient was denoted with a lowercase f, instead of a capital F. We have changed this as well.

```r
# Remove extra spaces and standardize to uppercase
diabetes_df$CLASS <- toupper(trimws(diabetes_df$CLASS))
diabetes_df$Gender[diabetes_df$Gender == "f"] <- "F"

diabetes_df |> count(CLASS)
```

```
  CLASS   n
1     N 103
2     P  53
3     Y 844
```

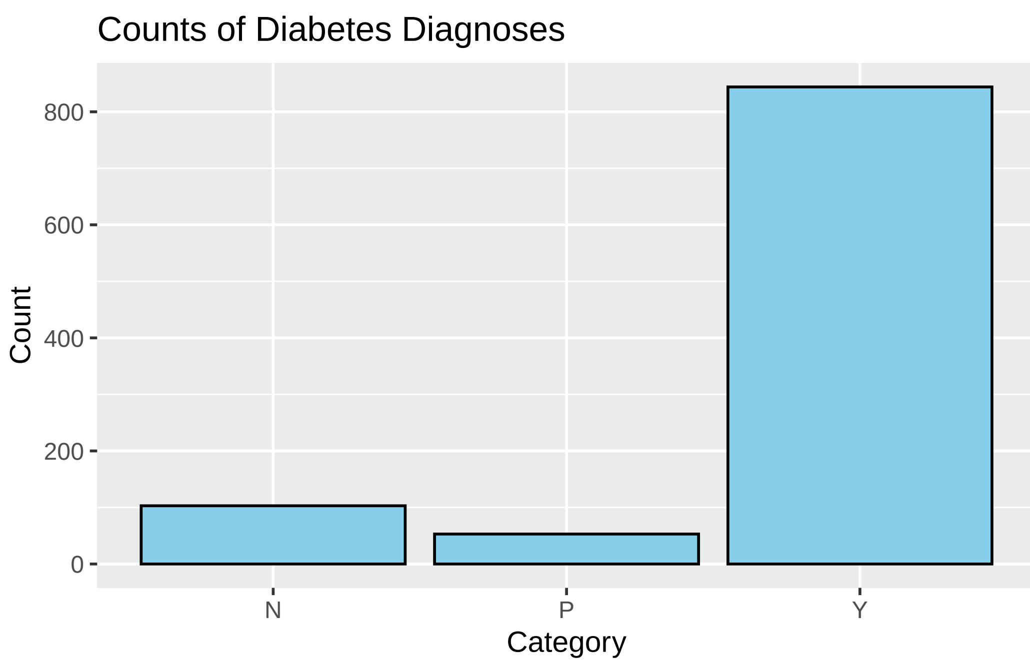
```r
diabetes_df |> count(Gender)
```

```
  Gender   n
1      F 435
2      M 565
```

We checked the counts again, and all seems to be consistent!

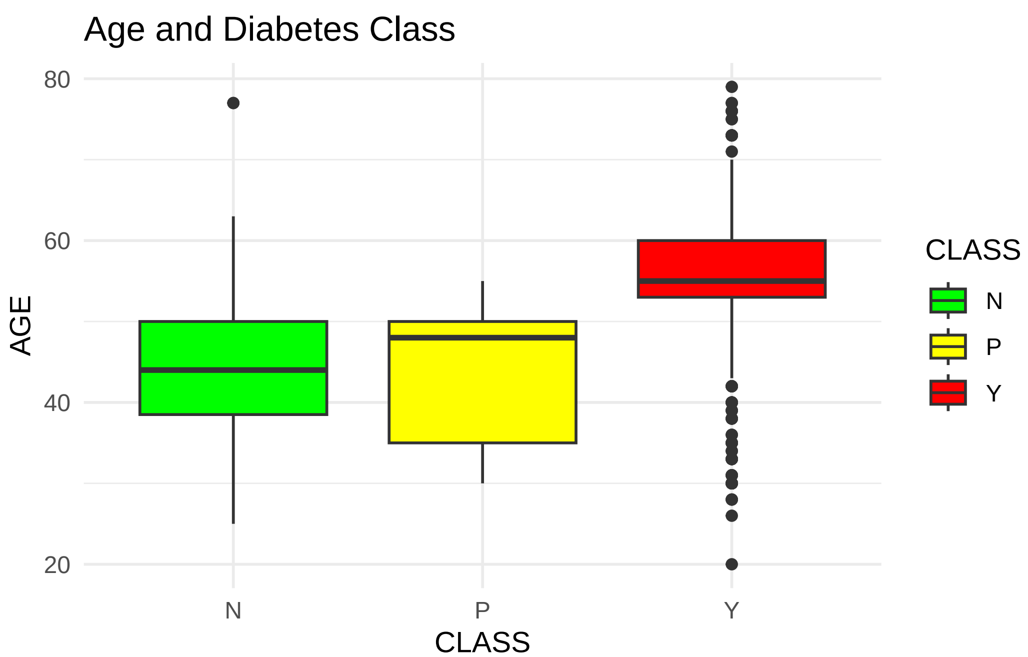
Demographics Plots

```
ggplot(diabetes_df, aes(x = CLASS)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(
    title = "Counts of Diabetes Diagnoses",
    x = "Category",
    y = "Count")
```

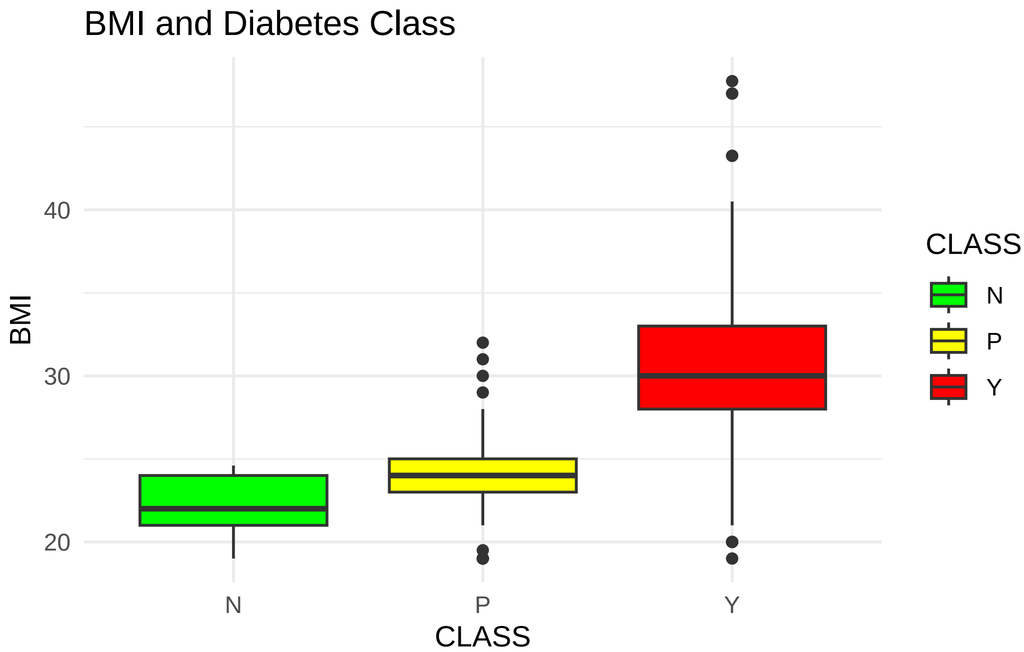


This bar graph shows the counts of patients who were identified as either non-diabetic (N), Pre-Diabetic (P), or Diabetic (Y). We can see that the vast majority of patients in the dataset were classified as diabetic, which could potentially influence analysis.

```r
ggplot(data = diabetes_df, aes(x = CLASS, y = AGE, fill = CLASS)) +
  geom_boxplot() +
  ggtitle("Age and Diabetes Class") +
  theme_minimal() +
  scale_fill_manual(values = c("Y" = "red", "P" = "yellow", "N" = "green"))
```




```r
ggplot(data = diabetes_df, aes(x = CLASS, y = BMI, fill = CLASS)) +
  geom_boxplot() +
  ggtitle("BMI and Diabetes Class") +
  theme_minimal() +
  scale_fill_manual(values = c("Y" = "red", "P" = "yellow", "N" = "green"))
```

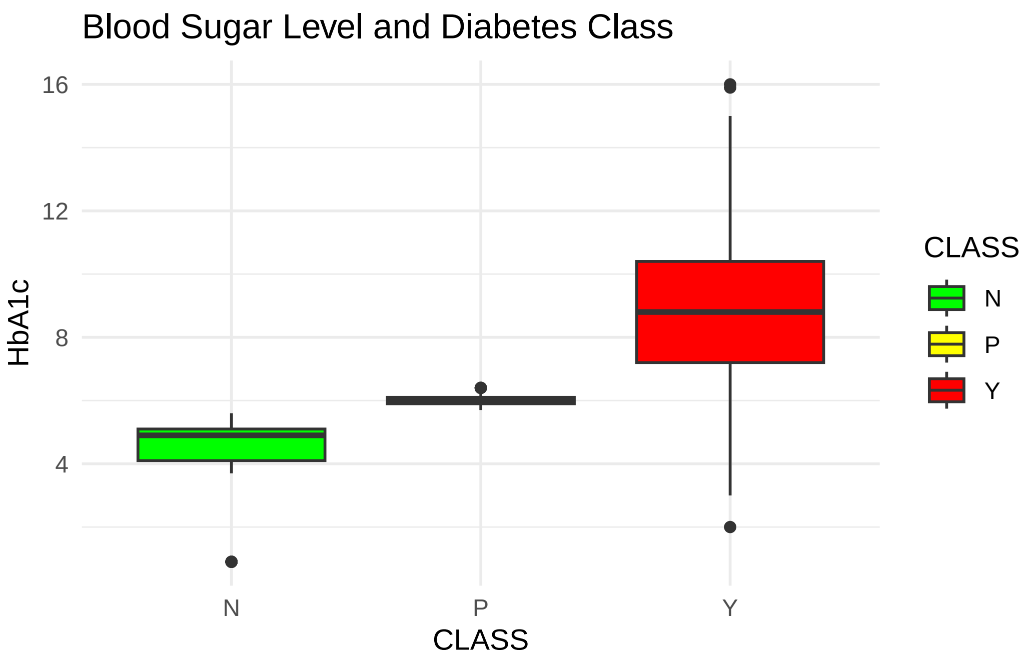


The first box plot illustrates the relationship between age and diabetes class (Y = Yes, P = Pre-diabetes, N = No). We can see that the median age increases progressively from the "No" (N) category to the "Pre-diabetes" (P) and then to the "Diabetes" (Y) category. This trend suggests that individuals with a diabetes diagnosis tend to be older on average compared to those without diabetes or in the pre-diabetes category. The distribution in the "Y" - or "Diabetes" - class shows a wider range of ages, indicating more variability in the ages of people diagnosed with diabetes.

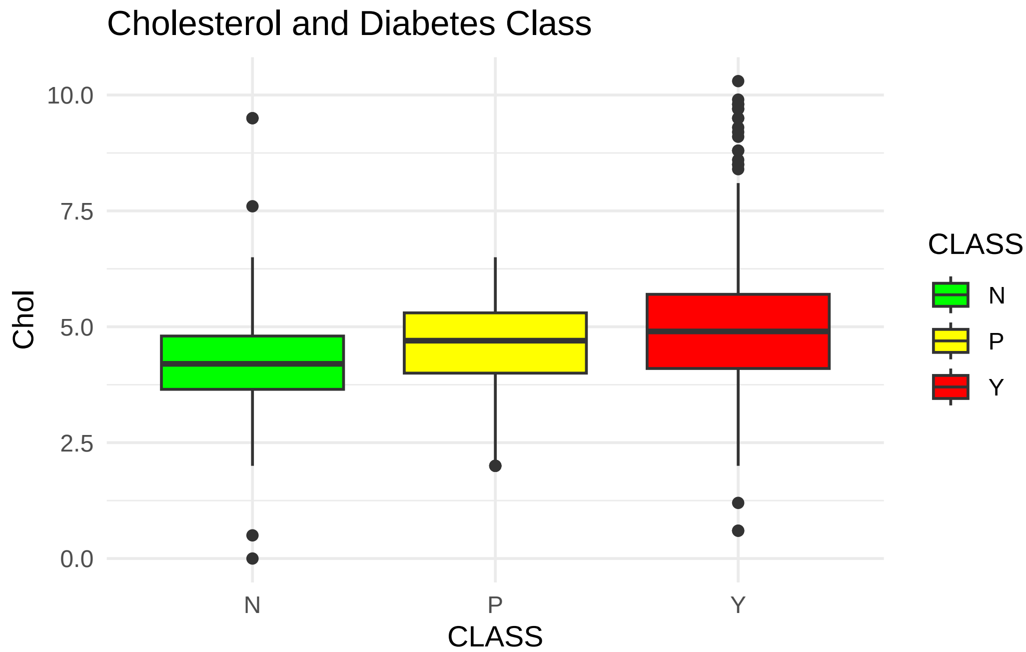
The second box plot shows the distribution of Body Mass Index (BMI) across the same diabetes categories. Here, we see a similar trend where the median BMI increases from the "No" (N) to "Pre-diabetes" (P) and then to the "Diabetes" (Y) category. Individuals with diabetes (Y) have a notably higher median BMI compared to those in the pre-diabetes or non-diabetes groups. The wider spread in the "Diabetes" category suggests that BMI varies more among individuals with diabetes, highlighting the potential link between obesity and the progression to diabetes.

Clinical Measures Plots

```
ggplot(data = diabetes_df, aes(x = CLASS, y = HbA1c, fill = CLASS)) +
  geom_boxplot() +
  ggtitle("Blood Sugar Level and Diabetes Class") +
  theme_minimal() +
  scale_fill_manual(values = c("Y" = "red", "P" = "yellow", "N" = "green"))
```




```
ggplot(data = diabetes_df, aes(x = CLASS, y = Chol, fill = CLASS)) +
  geom_boxplot() +
  ggtitle("Cholesterol and Diabetes Class") +
  theme_minimal() +
  scale_fill_manual(values = c("Y" = "red", "P" = "yellow", "N" = "green"))
```



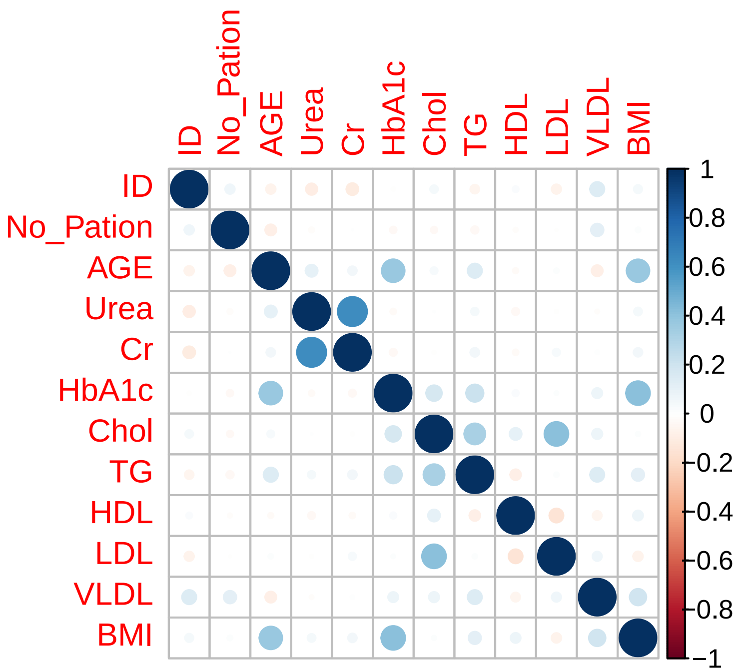
The first box plot shows the relationship between blood sugar levels (HbA1c) and diabetes class. It is evident that there is a dramatic difference in median HbA1c levels from the increases from the "Non-Diabetic" (N) category to the "Diabetes" (Y) category. This is expected, as medical knowledge states that individuals with higher blood sugar levels are more likely to be at risk for pre-diabetes or diabetes. The data indicates that blood sugar levels are higher and more variable among those diagnosed with diabetes.

In contrast, the second box plot shows the distribution of cholesterol levels across the same diabetes categories. Here, we do not observe a clear trend in cholesterol levels between the "No" (N), "Pre-diabetes" (P), and "Diabetes" (Y) groups. The median cholesterol levels remain relatively stable across all categories, with no significant increase as the diabetes class changes. This suggests that cholesterol may not be as closely related to the classification of diabetes as blood sugar levels (HbA1c), or that its role in diabetes risk may be more complex or influenced by other factors.

EDA Inquiry for Research Question 2: How correlated are the predictor variables (ex. blood sugar level, cholesterol) that may be utilized to predict the diabetes class of a patient?

- This may serve to better understand the multicollinearity included within the model, and how to remove variables to strengthen the statistical interpretation of the remaining predictor variables included within the model.

```
df_corr=cor(diabetes_df[sapply(diabetes_df,is.numeric)])
corrplot(df_corr)
```



From this correlation matrix of the numerical predictor variables, it appears that there will not be major issues of multicollinearity when preparing the model. The strongest correlation appears to be between age and BMI, or Urea and Creatinine level. One of either of these variables may be excluded from the regression, depending on testing.

Dataset 2

Data source:

https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Brief description:

This data set reflects incidents of crime in the City of Los Angeles dating from 2020 to 2023. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data.

This code book describes the data in more depth: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data

Long description

The City of Los Angeles provides a crime dataset from 2020 to the 2023, covering incidents reported to the Los Angeles Police Department (LAPD). Here's a breakdown of key columns in the dataset:

1. DR_NO: Unique identifier assigned to each crime report, used to track individual cases.

2. Date Rptd: Date the incident was officially reported to the police.

3. DATE OCC: Date when the crime actually occurred.

4. TIME OCC: Time the incident occurred, which allows for time-of-day analysis.

5. AREA: Code representing the geographical area of Los Angeles where the incident took place.

6. AREA NAME: Name of the area corresponding to the AREA code, providing a more human-readable location.

7. Rpt Dist No: Reporting district number within the LAPD, which is a more specific geographical indicator within the area.

8. Part 1-2: Crime classification indicator distinguishing between Part 1 and Part 2 crimes, which helps in crime severity categorization.

9. Crm Cd: Crime code representing a specific type of crime.

10. Crm Cd Desc: Description of the crime type (e.g., robbery, assault), giving context to the "Crm Cd" column.

11. Mocodes: Modus operandi codes that describe the method or behavior pattern of the suspect.

12. Vict Age: Age of the victim, which can be used for demographic analysis.

13. Vict Sex: Gender of the victim, another demographic detail.

14. Vict Descent: Ethnic background or descent of the victim.

15. Premis Cd: Premise code indicating the type of location where the crime occurred.

16. Premis Desc: Description of the premise (e.g., street, residence), helping to understand crime locations.

17. Weapon Used Cd: Code indicating if a weapon was used in the crime, which can be used to assess weapon involvement trends.

18. Weapon Desc: Description of the weapon, if applicable, providing details on the weapon type.

19. Status: Code indicating the current status of the case (e.g., open, closed).

20. Status Desc: Description of the case status, complementing the “Status” code with a text explanation.

21. Crm Cd 1-4: Additional crime codes, capturing cases where multiple types of crimes were involved in a single incident. 22. LOCATION: General location description of the crime.

23. Cross Street: Cross street information for more precise location data. 24. LAT: Latitude coordinate of the crime location, useful for mapping. 25. LON: Longitude coordinate of the crime location, also useful for mapping.

This dataset allows for comprehensive analysis of crime trends in Los Angeles, with potential insights into crime types, locations, times, demographics of victims, weapon involvement, and case status. The dataset is valuable for identifying patterns, conducting demographic analysis, and mapping geographical hotspots of crime.

Research question 1:

Research Question: What is the relationship between the severity of reported crimes and their spatiotemporal distribution in L.A? More specifically how do the frequencies of Part 1 and Part 2 violent crimes vary across different geographical areas of Los Angeles and the time of the day between 2020 and 2023?

Outcome Variable: *Part 1-2* (binary variable indicating crime seriousness, with Part 1 crimes generally more serious than Part 2).

Predictor Variables: *LAT* and *LON* (location coordinates), *AREA* (area code), and *TIME OCC* (time of day). Maybe others too.

Inference Goal: This question seeks to determine if more serious crimes are more prevalent in certain areas and at specific times.

Research question 2:

Research Question: How do victim demographics (age, sex, descent) influence the likelihood of being involved in the most common crimes—vehicle theft, simple assault (battery), burglary from vehicle, theft of identity, and felony vandalism—in Los Angeles between 2020 and 2023, and how do these patterns vary across different geographical areas?

Outcome Variable Name: `Crm.Cd.Desc` (Crime Code Description)
Type: Categorical Variable (Nominal)
Description:
The outcome variable for this research question is the **type of crime committed**, specifically focusing on the top five most common crimes in Los Angeles between 2020 and 2023. These crimes are:

1. **Vehicle - Stolen**
2. **Battery - Simple Assault**
3. **Burglary From Vehicle**
4. **Theft of Identity**
5. **Vandalism - Felony ($400 & Over, All Church Vandalisms)**

This variable represents the specific crime associated with each incident report in the dataset. It is a nominal categorical variable because the categories (crime types) are names without an inherent order or ranking.

Predictor Variables: - Vict.Age - Vict.Sex - Vict.Descent - AREA.NAME

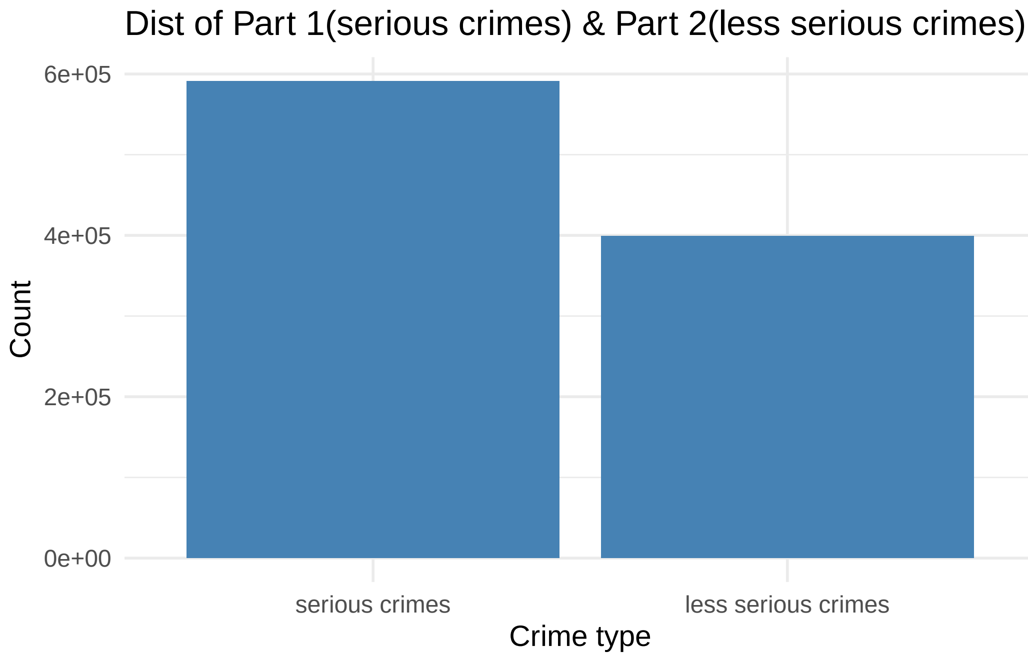
Inference Goal: This question aims to analyze the relationship between victim demographics and the likelihood of being involved in the five most common crimes in Los Angeles. It also seeks to investigate how these patterns differ across various geographical areas within the city. By focusing on these specific crimes and demographic factors, we can identify potential vulnerabilities and trends in victimization across different population groups and locations.

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

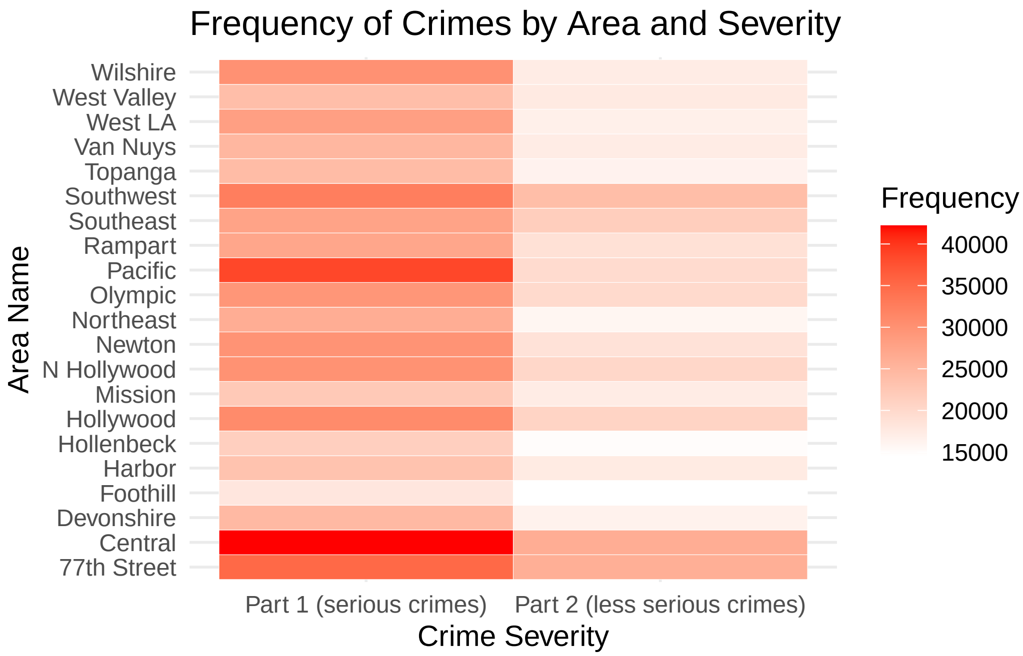
```r
data <- read.csv("https://github.com/lilah-duboff/Stats_Final_Project/raw/refs/heads/main/Cri

glimpse(data)
```

```
Rows: 990,293
Columns: 28
$ DR_NO          <int> 190326475, 200106753, 200320258, 200907217, 220614831, ~
$ Date.Rptd      <chr> "03/01/2020 12:00:00 AM", "02/09/2020 12:00:00 AM", "11~
$ DATE.OCC       <chr> "03/01/2020 12:00:00 AM", "02/08/2020 12:00:00 AM", "11~
$ TIME.OCC       <int> 2130, 1800, 1700, 2037, 1200, 2300, 900, 1110, 1400, 12~
$ AREA           <int> 7, 1, 3, 9, 6, 18, 1, 3, 13, 19, 18, 19, 2, 10, 3, 18, ~
$ AREA.NAME      <chr> "Wilshire", "Central", "Southwest", "Van Nuys", "Hollyw~
$ Rpt.Dist.No    <int> 784, 182, 356, 964, 666, 1826, 182, 303, 1375, 1974, 18~
$ Part.1.2       <int> 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2~
$ Crm.Cd         <int> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
$ Crm.Cd.Desc    <chr> "VEHICLE - STOLEN", "BURGLARY FROM VEHICLE", "BIKE - ST~
$ Mocodes        <chr> "", "1822 1402 0344", "0344 1251", "0325 1501", "1822 1~
$ Vict.Age       <int> 0, 47, 19, 19, 28, 41, 25, 27, 24, 26, 26, 8, 7, 0, 56,~
$ Vict.Sex       <chr> "M", "M", "X", "M", "M", "M", "M", "F", "F", "M", "M", ~
$ Vict.Descent   <chr> "O", "O", "X", "O", "H", "H", "H", "B", "B", "H", "B", ~
$ Premis.Cd      <int> 101, 128, 502, 405, 102, 501, 502, 248, 750, 502, 501, ~
$ Premis.Desc    <chr> "STREET", "BUS STOP/LAYOVER (ALSO QUERY 124)", "MULTI-U~
$ Weapon.Used.Cd <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, 400, NA, 400, 400, ~
$ Weapon.Desc    <chr> "", "", "", "", "", "", "", "", "", "STRONG-ARM (HANDS,~
$ Status         <chr> "AA", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "IC", "~
$ Status.Desc    <chr> "Adult Arrest", "Invest Cont", "Invest Cont", "Invest C~
$ Crm.Cd.1       <int> 510, 330, 480, 343, 354, 354, 354, 354, 354, 624, 354, ~
$ Crm.Cd.2       <int> 998, 998, NA, NA, NA, NA, NA, NA, NA, NA, NA, 821, 860,~
$ Crm.Cd.3       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Crm.Cd.4       <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ LOCATION       <chr> "1900 S  LONGWOOD                     AV", "1000 S  FLO~
$ Cross.Street   <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "VA~
$ LAT            <dbl> 34.0375, 34.0444, 34.0210, 34.1576, 34.0944, 33.9467, 3~
$ LON            <dbl> -118.3506, -118.2628, -118.3002, -118.4387, -118.3277, ~
```

Exploratory Plots: ***Research question 1 plots:***

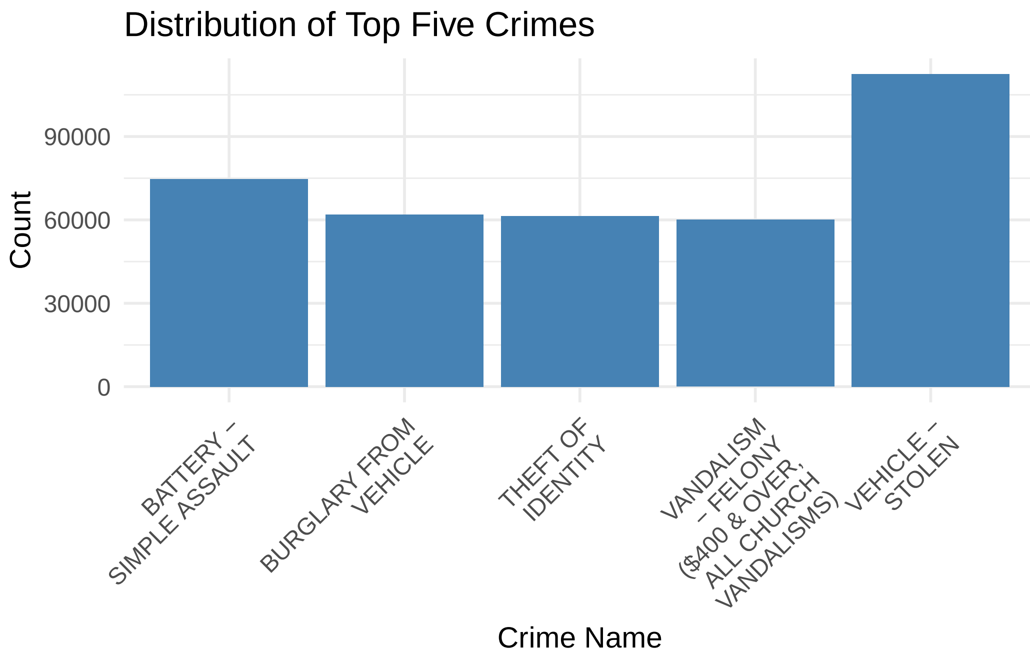


Distribution of Part 1 (Serious Crimes) & Part 2 (Less Serious Crimes) This bar chart categorizes crimes into Part 1 (serious crimes) and Part 2 (less serious crimes) and shows the frequency of each category. The two bars illustrate the relative proportions of serious versus less serious crimes, providing insight into the overall severity distribution within the dataset. EDA Insights: Part 1 crimes have a higher frequency than Part 2 crimes, indicating that serious crimes make up a larger portion of the reported incidents. The more severe crimes have about 591,254 observations and less severe crimes are 399,039. This distribution helps understand the nature of crime severity in Los Angeles, which may influence policing or resource allocation.



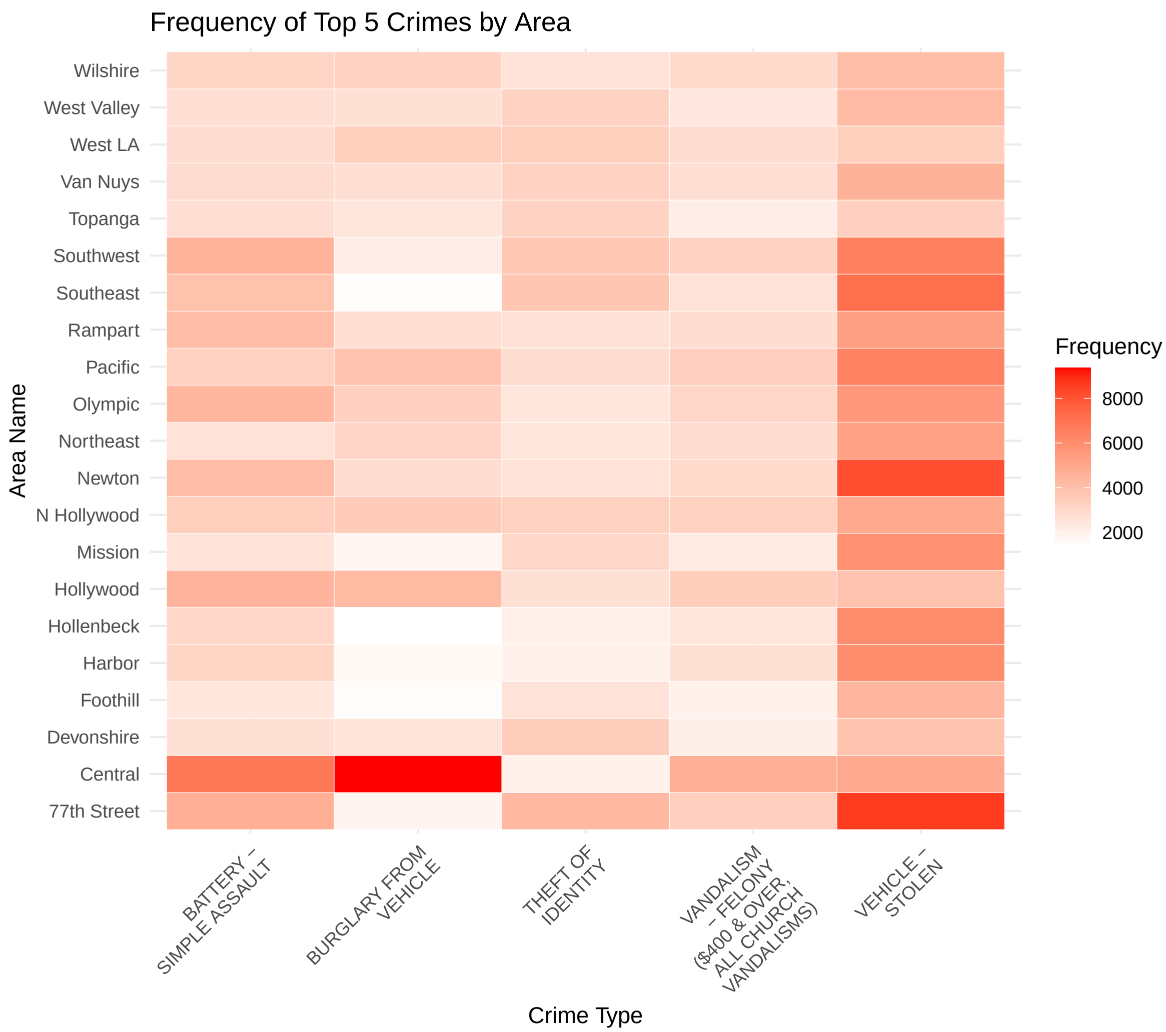
Frequency of Crime Severity(more severe and less severe) Description: This heatmap shows the frequency of the top five crimes across various geographical areas. Each cell's color intensity represents the count of a particular crime type within an area, with darker shades indicating higher frequencies. The y-axis lists the areas, while the x-axis lists the severity of crime(binary: less severe and more severe), allowing for a spatial view of crime distribution. EDA Insights: Areas with darker cells have higher crime counts, highlighting regions with potentially higher crime rates for each type. Certain areas like "Pacific", "Central","77 Street", etc. have a consistently high frequency for more severe crimes, indicating general high-crime zones. Areas with high frequencies for severity of the crime could benefit from targeted crime prevention programs or policing strategies.

Research question 2 plots:

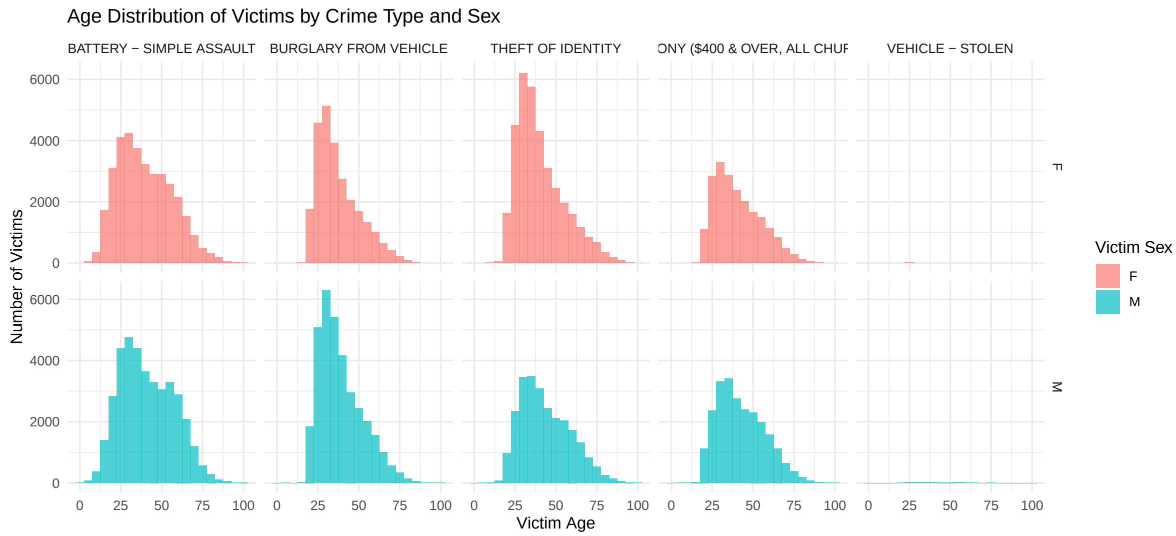


Distribution of Top Five Crimes Description: This bar chart shows the frequency of the top five most common crimes in Los Angeles. Each bar represents a specific crime type, with the height indicating the total count of occurrences. This visualization provides an overview of the relative prevalence of each crime, helping to identify which types are most frequent in the dataset. EDA Insights: The chart clearly shows that "Vehicle - Stolen" has the highest frequency among the top five crimes, followed closely by "Battery - Simple Assault." This distribution allows for quick comparison across crime types, highlighting which crimes are more common. Observing high counts for specific crimes could indicate priority areas for law enforcement or community awareness programs.

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



Frequency of Top 5 Crimes by Area Description: This heatmap shows the frequency of the top five crimes across various geographical areas. Each cell's color intensity represents the count of a particular crime type within an area, with darker shades indicating higher frequencies. The y-axis lists the areas, while the x-axis lists the crime types, allowing for a spatial view of crime distribution. EDA Insights: Areas with darker cells have higher crime counts, highlighting regions with potentially higher crime rates for each type. Certain areas like "Pacific", "Central","77 Street", etc. have a consistently high frequency for crime types like "Vandalism" and "Central" area has highest " Burglary from Vehicle" crime type cases. Areas with high frequencies for specific crime types could benefit from targeted crime prevention programs or policing strategies.



Age Distribution of Victims by Crime Type and Sex Description: This faceted plot shows the age distribution of victims across different crime types, separated by victim sex (female and male). Each facet represents a unique combination of crime type and sex, with the x-axis showing victim age and the y-axis showing the count. This plot reveals age-based patterns in victimization for each crime type, broken down by gender. EDA Insights: Certain age groups may have higher victimization rates for specific crimes, suggesting patterns of vulnerability related to age. Comparing male and female distributions within each crime type can reveal gender-based differences in victimization, potentially indicating targeted or biased victimization. There is a noticeably higher count of female victims across multiple crime types, particularly in simple assault and identity theft. Male victim counts are generally lower than females for most crimes; however, certain crimes, such as vehicle theft and burglary from vehicles, show a more balanced distribution between genders. This could suggest that these crime types are less influenced by the victim's gender. Across most crimes, young adults appear to have the highest victimization rates, particularly visible in crimes like simple assault and identity theft. This age group might be more exposed to environments or activities associated with these crimes. Different crimes show unique distributions by age and gender, which can guide more focused safety or awareness campaigns for specific demographic groups.