This EDA assignment is intended to get you interacting with and thinking about your final project data. Simply looking at the data—remember, humans are visual creatures—is sometimes enough to understand how variables relate (inference) and guide your research agenda.

No need to show your code on this assignment.

Format [20 points]

Breakdown

Content	Full Points Points	Full Points Desc.	Half Points Points	Half Points Desc.	No Points Points	No Points Desc.
Section headers	5	1 per header				
Plots						
Labelled axis	8	1 per label				
Titles	4	1 per title				
Set text size to 20	3	1 per plot				
Typos/grammar errors	10	none to some	5	many	0	unintelligible

Create a header in your document for the following sections:

Introduction [5 points]

Provide a paragraph of containing your prediction problem, motivation, and context in whichever order suits you best. Improve upon what you wrote for the proposal. Do not copy and paste from proposal; we can always improve our writing. You will need to have compelling writing in your career. This is good practice.

Data Wrangling and Cleaning [10 points]

Provide the name of the data set(s) and the source from where your data are retrieved. Describe (don't show) what data wrangling and cleaning was necessary to create your final data set.

Response Variable [20 points]

Breakdown

Regression	Classification

Points	Description	Points	Description
4	Explain transformation or not of y	10	Percent of observations each class contains in a table
6	Provide mean, median, and sd of final y		
10	Histogram of final y	10	Barplot of y

Covariates [15 points]

Breakdown

Points	Description
5	Describe transformation of covariates (if any)
5	Describe interactions of covariates (if any)
5	Explanation of why transformations and interactions were made (or not if none)

(gg)Plots [30 points]

In this section you will create plots of y with your expected three most important covariates from your proposal. The covariates must match your proposal variables. Please contact instructor for extenuating circumstances if this criteria cannot be met.

5 points are awarded for creating each plot.
5 points are awarded for a correct written interpretation of the relationship between the plotted variables for each plot.

Continuous vs. Continuous Variables: Use either binned scatter plots with 100 bins or hist2d plots, whichever better shows the correlation (or lack thereof).

Continuous vs. Discrete Variables: Use violin plots with the x-axis set to the discrete variable.

Discrete vs. Discrete Variables: Use tile plots using the following code skeleton framework:

```
df = table(name_of_your_data[, c('x', 'y')])
ggplot(df, aes(x = x, y = y, fill = Freq)) +
  geom_tile()
```