````
---
title: "Station 2: EDA with R"
author: "Your name here"
output: html_document
editor_options:
  chunk_output_type: console
---

A. Introduction
-----------------

In this document you will be introduced to the functions we will use most frequently in this class.  If you are moderate to advanced R user, feel free to use whatever functions you'd like to accomplish the tasks. The first part of this document contains explanations for beginning R users, but all students should work through the entire document.

**TASK--Run the code below to set options for the document and load packages.**


```{r, include = F}
##  NOTE - in all future documents/assignments this code will be included for you and you are expected to run it without prompting.

# Clear Workspace
rm(list = ls())

# load packages we typically use for this class.
library(mosaic)
library(ggformula)
library(Stat2Data)
library(tidyverse)
```

B. Data
-----------------
## Load data

Doing statistics requires *data*.  R comes with many built-in datasets, and the packages you loaded above (specifically Stat2Data) give you access to even more.  To load a dataset from within R you can use the `data` command:

**TASK--Use the code below to load the dataset diamonds, which is part of the ggplot2 package and included in the tidyverse.**

```{r}
data("diamonds")
```

This is a great chance to remind you that R cares about letter case.  This means that data(diamonds) and data(Diamonds) actually load different datasets!  In this case, please make sure to load the *lowercase letter* `diamonds` dataset.

### Inspecting the data source

Now you're ready to learn a little bit about the `diamonds` data set.

```{r}
# Inspecting the data source
glimpse(diamonds)
head(diamonds)
names(diamonds)
```

**TASK:** Edit the bullet list below to add a short description in your own words describing
````

what each function does.

- `glimpse()` : this function...
- `head()` : this function...
- `names()` : this function...

Some Data Prep

The following is a little bit of data wrangling to get the source data in shape for our purposes. You can ignore this part for now, and we can talk about it another time.

```{r}
# Recode & filter (no edits needed - just run this chunk)
recoded <- # make a new dataset called recoded
  diamonds %>% # by starting with the diamonds data
  filter(color=="D" | color=="J") %>% # and filtering observations to keep only colors D and J
  mutate(col = as.character(color)) # tell R some specifics about how to record the variable color.
```

Basically, we're going to do a bit of exploration of variables that impact cost of diamonds. Even if you haven't used R before, you might be able to tell from the code that we start with the diamonds data, filtered (i.e. restricted) our data set to only include the diamonds that are either color "D" or "J", created a new categorical variable called "col", and stored the whole thing in a new data set called "recoded".

Statisticians should always know something about the data domain in order to be useful. If you don't know anything about the subject area you need to at least learn some basics. Wikipedia is usually a good place to start: https://en.wikipedia.org/wiki/Diamond_color.

C. Exploratory Data Analysis

For the purposes of our class, it's useful to learn a model-centric approach to R. The psuedo-code below is going to be our foundation for the rest of the class:

`function( Y ~ X, data = DataSetName )`

Here's a short description of each part in the pseudo-code above:

- `function` is an R function that dictates something you want to do with your data; for example,
    - `mean` calculates the mean
    - `t.test` performs a t-test
    - `lm` fits a linear regression model
- `Y` is the outcome of interest (response variable)
- `X` is some explanatory variable; you can use "1" as a placeholder if there is no explanatory variable
- `DataSetName` is the name of a data set loaded into the R environment

Always start with clear research questions. Our question for this exercise: *How do diamond prices compare for "D" and "J" colored diamonds?*

The purpose of the exploratory data analysis (EDA) is to learn as much as you can about your research question **before doing any fancy statistical modeling**. We basically want to try and answer the research question with EDA if possible...or at least have a guess as to what the answer "should" be. Then we use statistical models to formally accommodate variability in the data and calculate the uncertainty of our conclusions.

Mean price by color

Use the R code chunk to calculate the mean price by color. Summarize your observations below the code chunk. Don't forget, we did some data wrangling above and made a new data set called "recoded." Use the "recoded" data for the rest of this analysis.

```{r}
mean(price ~ col, data = recoded)
```

TASK--Share your observations:

Other summary statistics by color

Of course, there are lots of other ways to summarize a numerical variable besides the mean. Use the R code chunk to calculate the other summary statistics for price of each diamond color using `favstats()`. Summarize your observations below the code chunk. How do the prices compare between D diamonds and J diamonds?

TASK--Produce the required code:

(Hint: Do you not know how `favstats` works? Well, it's a function just like any other: it follows the syntax described at the top of this section! Also, you can always find details and examples by searching the Help menu in the lower-right quadrant.)

```{r}

```

TASK--Summarize your observations:

2-3 Basic plots of the data

Make side-by-side boxplots along with a scatterplot using the R code shown below.

TASK--Run the code to make a boxplot; then modify the 2nd bit of code to make the desired scatterplot:

```{r}
# make a boxplot of price by color
gf_boxplot(price ~ col, data = recoded)

# make a scatter plot of price versus carat
#gf_point(response ~ explanatory, data = recoded)

```

TASK--Share your observations:

Multivariable relationships

The world is often too complicated to be understood by studying one or two variables at a time. Color is certainly not the only variable that impacts the value of a diamond. You may have noticed in the Wikipedia article that color is only one of the "4 C's" that influence the value of a diamond.

Adjusting for other variables

Create a scatter plot of price vs carat (a measure of diamond weight) that colors each plotted point according the color of the diamond it represents. Write a few sentences below the code chunk to explain what you've observed from this plot.

TASK--modify the code to produce the scatterplot:

```{r}
#gf_point(response ~ explanatory1, color = ~ explanatory2, data = recoded)
```

TASK--Share your observations:

D. US States Data Analysis

Use data from an outside source

It's important to be able to download a file from an outside source (typically Moodle) and then load it in R for use. There are multiple ways to do this, but one is described in Station_1. If you don't know how to do this, see Station_1 for instructions, or ask your partner.

TASK--Download the file US_States_fall2021.csv from Moodle and load the data into your environment

Once the data is loaded into the Environment, that means it's loaded in the Console, NOT this RMarkdown document! Make sure you copy the code that loaded it into the Console into the R chunk below. Call the data set `state_data`.

```{r}

```

TASK--Use head() and glimpse() to explore the format of the data a bit:

```{r}

```

This data has one row for each of the states, and each variable is recorded for the entire state. For example, that variable 'HeavyDrinkers' gives the percentage of residents in a state that are heavy drinkers. According to these data 5.9% of Pennsylvania residents are heavy drinkers.

TASK--Refer to the US States Merged Data codebook (also on Moodle) and select some variables of interest:

You can choose whatever variables you want. You will explore these variables (and relationships between them) in Stations 2, 3, and Homework 1.

- Quantitative variable 1: ...
- Quantitative variable 2: ...
- Categorical variable 1: ...
- Categorical variable 2: ...

EDA for State data

For all the tasks below, you will have to alter the code given for *your* choosen variables!

TASK--Determine how many states are in each category of a categorical variable:

```{r}
#tally( ~ CatVariable1, data = state_data)
```

TASK--create a barchart for one categorical variable:

```{r}
#gf_bar( ~ CatVariable1, data = state_data)
```

TASK--Determine how many states are in combinations of categories for two categorical variables:

```{r}
#tally(CatVariable1 ~ CatVariable2, data = state_data)
```

TASK--create a side-by-side barchart for two categorical variables:

```{r}
#gf_bar( ~ CatVariable1, fill = ~ CatVariable2, data = state_data, position = position_dodge())
```

TASK--create a histogram for one of your quantitative variables:
```{r}
#gf_histogram( ~ QuantVariable1, data = state_data)
```

TASK--create side-by-side boxplots for one quantitative and one categorical variable:

```{r}
#gf_boxplot(QuantVariable1 ~ CatVariable1, data = state_data)
```

TASK--create a scatterplot for two quantitative variables:

```{r}
#gf_point(QuantVariable2 ~ QuantVariable1, data = state_data)
```

TASK--create a scatterplot for two quantitative variables, using a categorical variable for color:

```{r}
#gf_point(QuantVariable2 ~ QuantVariable1, color = ~ CatVariable1, data = state_data)
```

Summarizing what you learned
Whenever you summarize the results of an analysis, it should always be **in context**. "In context" means that you talk about variables, not codes or shorthands like "CatVariable1". A good question to ask yourself is: "If I walked up to a person on the street and told them this fact about my data, would they understand what was talking about?"

Example: BAD

You: "Hi, did you know that there is a positive relationship between QuantVariable1 and QuantVariable2? It looks like the r is pretty strong."

Person on street: "WTF?"

Example: GOOD

You: "Hi, did you know that states with higher 8th grade math test scores also have higher average IQ's? The relationship between these two variables is quite strong!"

Person on street: "How interesting, although maybe not surprising? Let's have a discussion about the difficulties of measuring 'intelligence' with either standardized testing or IQ tests. What is 'intelligence', anyway?"

TASK--share three interesting observations you learned from these figures (in Section D). Anything you learned must be in context!

Note: Make sure you save your work. You will need it for your first homework assignment.