

# Mini Project 2: Exploring Data

**Due:** Oct 17 by 11:59pm

**Weight:** This assignment is worth 10% of your final grade.

**Purpose:** When you get a data file to analyze, one of the first things you'll want to do is *explore* it. But that's often easier said than done. The purpose of this assignment is to give you a chance to practice some of the techniques we've discussed in class for exploring a data set.

**Assessment:** Your submission will be assessed using the [rubric](#) at the bottom of this page.

## 1. Get organized

Download and **unzip** [this template](#) for your project, then open the `report.Rproj` file.

Once RStudio opens, click on the `report.qmd` file. That is the primary file you will edit to conduct your analysis.

## 2. Load the data

For this assignment, you'll be exploring data on the costs of hundreds of transit projects around the world collected by the [Transit Costs Project](#). The data span more than 50 countries and totals more than 11,000 km of urban rail built since the late 1990s.

Write code to load the `transit_cost.csv` file in the "data" folder.

## 3. Document the data

**Format:** This can just be a single paragraph describing the data sources.

Go to [this GitHub page](#) and read about the data we'll be using (that's where I got the data in from). In your `report.qmd` file, document the original data source as well as the source from where you downloaded the data. Note that the GitHub page is **not** the original data source - it is just a repository where someone else put the data.

## 4. Preview the data

**Format:** This can be a mix of code chunks and written text. It is more for you than me - think of this like jotting down notes to help you understand what is in the data.

Preview the data (e.g. using `head()`, `glimpse()`, `View()`, and / or make some quick plots). Take note of what variables are available, their types, and what they measure (Hint: look at the data dictionary on the GitHub page!).

## 5. Identify research questions

**Format:** Use a numbered list to list each question and the associated variables you will explore.

Once you have a sense for the available variables in the data, list **at least three** questions you think you may be able to answer with these data (you can list more if you want). For each question, also note which variables you plan to explore to address the question. A question can be about what the data captures (e.g. “Which countries in the data have the longest total amount of projects in km?”) or about a relationship between different variables (e.g. “Are projects with tunnels more expensive on a \$/km basis than projects without tunnels?”).

**Note:** It is okay if you end up not being able to answer your question - just write down what you think you *might* be able to find out by exploring the data.

## 6. Explore the data

Go through each of your questions and search for answers. **You should include at least one visualization for each question** (so a minimum of 3 charts total).

For each question, follow these steps:

1. If necessary, modify the `transit_cost` data frame to address your question. For example, you may need to create new variables (e.g. using `mutate()`), filter the data (e.g. if you're only interested in a single country, year, etc.), or you may want to rename some of the variables. For this task, you'll be relying on your data wrangling skills.
2. Examine summary measures (centrality, variability, correlation) in the variables relevant to your question by making charts and / or printing out summary values / tables. Your chart(s) should be appropriately chosen according to the data type and / or relationship you are searching for. Your charts should follow the design principles we have covered in class. They do not have to be fully “polished” yet, but at a minimum they should be accurate (i.e. not misleading) and they should not include distracting non-data ink.
3. For each research question, write at least one paragraph describing what you found. If you found an answer to your question, make sure you have included at least one chart (or a few charts if appropriate)

that helps explain your answer. If this process did **not** lead to an answer your question, write about how you might adjust your question or perhaps what other data you may need to address your question.

## 7. Render and submit

Click the “Render” button to compile your `.qmd` file into a html web page. Then open the `report.html` file in a web browser and proofread your report.

Does all of the formatting look correct? **Make sure there are no errors in the rendered file before submitting it.**

Once you’ve proofread your report, create a zip file of all the files in your R project folder for this assignment and submit it on the corresponding assignment submission on Blackboard.

## Bonus (+3%)







One logical relationship you might expect in this data is a strong correlation between the length of a transit line and it’s cost. But if we plot a scatterplot of `length` and `real_cost` (making sure to convert `real_cost` to a number), the correlation doesn’t look too strong. That’s because length and cost don’t scale together in linear space - they scale in log-linear space! That is, the *log* of these variables correlate much more strongly than the variables themselves. To get the bonus credit, make a plot that shows this relationship (hint: read about [log-log plots](#)).

## Grading Rubric

### 45 Total Points

Category	Excellent	Good	Needs work
Formatting	<b>5</b> Followed all formatting guidelines.	<b>4</b> Followed most formatting guidelines.	<b>3</b> Missing multiple formatting guidelines.
Documentation	<b>5</b> Original and downloaded data sources clearly described.	<b>4</b> Poor description of original and / or downloaded data sources.	<b>3</b> Poor / unclear / missing description of either original or downloaded data source
Research questions	<b>10 / 9</b> At least 3 RQs & associated data variables listed.	<b>8 / 7 / 6</b> <3 RQs listed or variables are missing.	<b>5 / 4 / 3</b> <3 RQs listed and variables are missing.
Exploration	<b>10 / 9</b> Summary measures and charts appropriately used to address all RQs; excellent summary description.	<b>8 / 7 / 6</b> Measures and charts used to address RQs could be improved; adequate summary description.	<b>5 / 4 / 3</b> Missing summary measures and / or charts to address RQs; poor summary description.

Category	Excellent	Good	Needs work
Visualizations	<b>10 / 9</b> Visualizations appropriately chosen according to data types and / or relationships.	<b>8 / 7 / 6</b> Visualizations are appropriate but could be improved.	<b>5 / 4 / 3</b> Poor match between visualization and data types / relationships, and / or missing.
Technical things	<b>5</b> All code runs without errors; all files included in the submitted .zip file.	<b>4</b> Code has only one or two error, otherwise runs; all files included in the submitted .zip file.	<b>3</b> Code has multiple errors; submitted .zip file is missing components necessary to reproduce analysis.

© 2023 John Paul Helveston   
 Wednesdays |  12:45PM - 3:15PM EST |  
 Monroe Hall 114  
 Dr. John Paul Helveston |  jph@gwu.edu  
 | 