Data Scientist/Data Analyst STC: Applicant Exercise:

**Background:**

World Bank project managers (task team leaders or TTLs) regularly assess the types of problems that may arise during project implementation and the effect that those problems will have on achieving the objectives of the project. These potential problems have been grouped into eight broad categories of risks to project development objectives. TTLs record their assessments by indicating if a specific risk category is Low (L), Moderate (M), Substantial (S), or High (H) for a given project. Thus, if a particular problem is highly likely to arise and will impede the achievement of the objective of the project, the TTL will rate that project as high risk in the corresponding category. TTLs also give an “Overall” risk rating and can use an “Other” category if they choose.

**Exercise:**

Attached, you will find two csv files. The first includes data on assessments of risk on a subset of World Bank projects. The second file includes different attributes of the same subset of World Bank projects. Please use them to answer the following questions and support your answers with visualizations. Please submit your code along with your answers and data visualizations.

1. Is risk related to the region in which a project is being implemented? For which risk is the relationship the strongest? How would you visualize the relationship between risk and region?
2. Please choose one other project attribute and analyze and report on its relationship to risk.

**Data Documentation:**

*risk\_data.csv*:

* project\_id: Unique identifier for each project
* risk\_rating\_code: Broad category of risk that is being assessed. There are 10 categories including an “Other” category.
* risk\_rating: Assessment of the level of risk in that category. This can be Low (L), Moderate (M), Substantial (S), or High (H).
* risk\_rating\_sequence: This column indicates which risk assessment it is. The lower the number the earlier in the project life cycle. Negative numbers indicate that the risk assessment occurred while the project was still being prepared.

*project\_data.csv*

* prod\_line: This refers to the type of project.
* project\_id: Unique identifier for each project. This corresponds to project\_id in the risk data file.
* proj\_emrg\_recvry\_flg: Indicates if the project is in response to an emergency.
* fcs\_indicator: Indicates if the project is in a Fragile or Conflict Situation.
* tl: Unique identifier of the project manager
* tl\_since: Date the current project manager came on to the project
* project\_track: Project preparation processing track. Track 1 involves fewer steps.
* scale\_up: Indicates if this project is a scale up of a previous project in the same country/sector
* len\_instr\_type: Type of loan instrument.
* practice: Indicates which global practice is leading the project. There are 14 global practices, which are similar to the main sector in which a project is working.
* country: Country in which the project is being implemented.
* region: One of six World Bank Group designated regions in which the project is being implemented.
* net\_commit\_amt: Value of the World Bank loan(s) associated with the project in millions of USD.
* grant: Value of any World Bank grants associated with the project in millions of USD.